ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Convergent QSAR Models for the Prediction of Cruzain Inhibitors

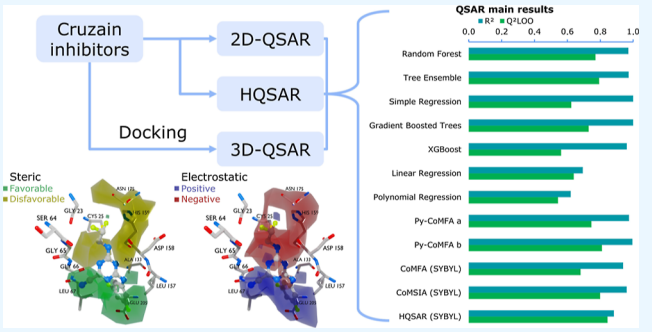
Rafael Bello Gonçalves,* Witor Ribeiro Ferraz, Raisa Ludmila Calil, Marcus Tullius Scotti, and Gustavo Henrique Goulart Trossini*

Cite This: *ACS Omega* 2023, 8, 38961−38982

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Chagas disease is a parasitosis caused by *Trypanosoma cruzi*. Cruzain, the major cysteine protease from *T. cruzi*, is an excellent therapeutic target in the search for antichagasic drugs. It is important in the role of cell invasion, replication, differentiation, and metabolism of the parasite. In this work, we developed and assessed multiple quantitative structure-activity relationship (QSAR) models for a set of 61 cruzain inhibitors. These models include two-dimensional (2D) QSAR, three-dimensional (3D) QSAR, such as comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA), and Hologram QSAR (HQSAR). In total, we generated 10 major and 114 minor model variations. Molecular docking was used to successfully align the molecules. All CoMFA and CoMSIA models, which incorporate multiple fields, demonstrated robustness in our analysis. Steric fields exhibited satisfactory convergence in the contour maps, while the electrostatic field converged into a single small region. The HQSAR model taking into consideration only Atoms and Connectivity, with fragment sizes ranging from two to five atoms, was considered the best of the HQSAR variations, despite exhibiting a higher level of deviance. In total, 78 model variations meet the minimum requirements to be considered acceptable. We found that using as few as five descriptors it is possible to obtain robust results with 2D-QSAR. Models such as Random Forest, Tree Ensemble, Linear Regression, and HQSAR are recommended for working with large data sets, while the 3D-QSAR models are intended to study the geometry of the ligands, to optimize them into new and better performing antichagasics. Virtual Screening of a set of hydrazones, guided by the top-performing models, identified promising candidates for experimental validation. Among them, dv007 and dv015 exhibited consistently high predicted pIC50 values (7.26 and 7.24, respectively), making them compelling candidates for further drug development.



1. INTRODUCTION

Chagas disease is a parasitic disease caused by *Trypanosoma cruzi* and one of the most prevalent neglected tropical diseases in Brazil. Endemic in 21 Latin American countries, it currently affects approximately six to seven million people worldwide, with an annual incidence of 30,000 new cases, and results in an average of 14,000 deaths per year and 8000 newborns infected during pregnancy. It is estimated that around 70 million people live in areas of exposure and are at risk of contracting the disease.[1]

There are only two drugs available (benznidazole and nifurtimox), both of which have serious adverse effects. The low efficacy of these drugs has worsened cases of resistance, leading to great concern and highlighting the need for the development of new chemotherapeutic agents.[2]

The selection of therapeutic targets is one of the key aspects of drug discovery in biochemical, selective, and safe processes. In this sense, cruzain, the main cysteine protease of *T. cruzi*, is essential in all evolutionary stages of the parasite, presenting itself as an interesting target in the search for new antichagasic agents. Another important point is that this enzyme is not found in humans, indicating selectivity and a reduced incidence of adverse effects.[3]

Several chemical classes have already been studied for their ability to inhibit cruzain, such as vinyl sulfones, triazoles, pyrimidines, thiosemicarbazones, chalcones, nitroalkenes, cyclic imides, and benzimidazoles.[4−9]

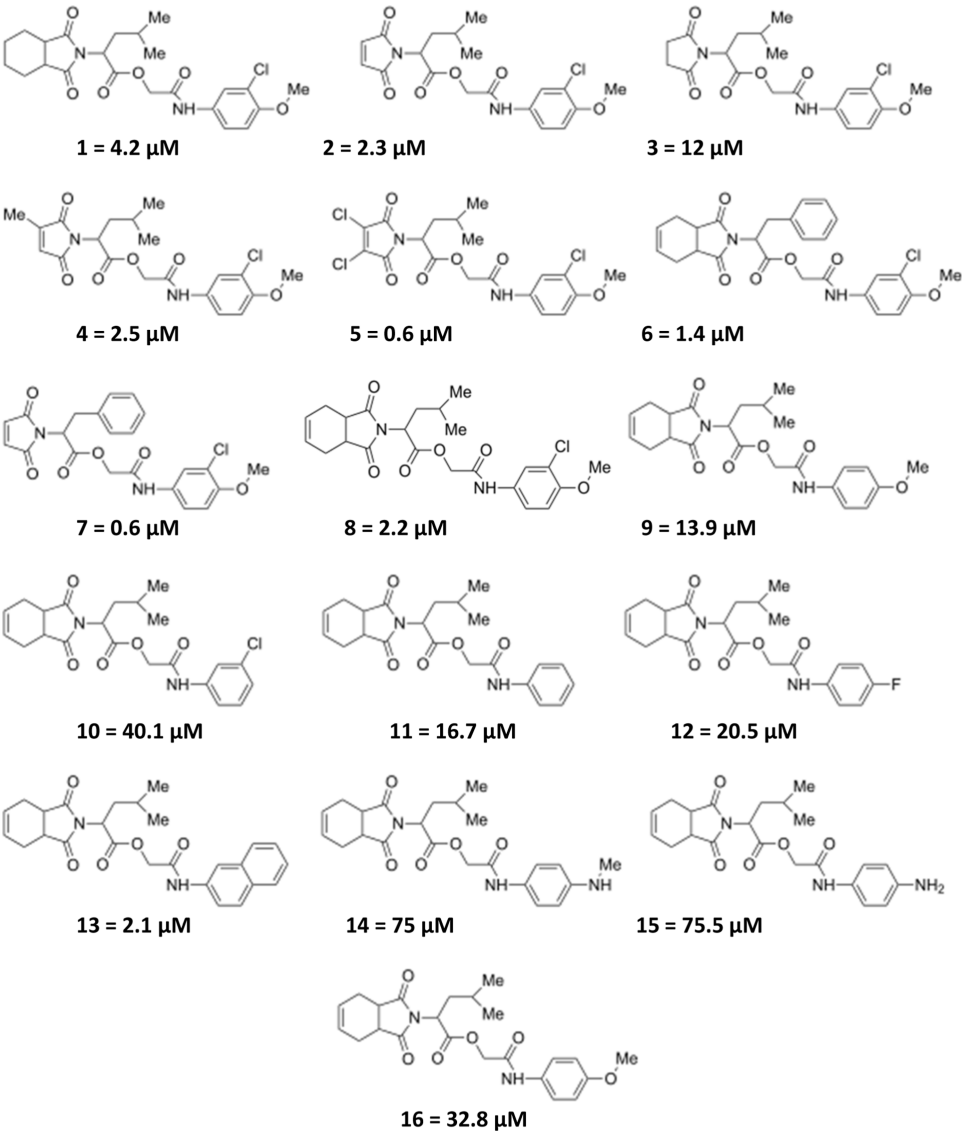
Prior research has investigated covalent inhibitors, including K777 (a vinyl sulfone derivative), and found that the irreversible binding of the ligand to the enzyme contributed to its high toxicity. Thus, most recent studies focus on the development of reversible inhibitors.[4]

Drug discovery benefits from an array of computational techniques. Molecular docking, one of those techniques, is a well-established and widely used method for drug design. It is capable of performing several tasks, such as molecular alignment,

Received: May 15, 2023
Accepted: September 29, 2023
Published: October 13, 2023

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Chart 1. First Set Is Composed Mostly of Cyclic Imides and Their Biological Activity (IC_{50})[4]



and this can be helpful to other tools, like the quantitative structure-activity relationship (QSAR), which has been used for decades to obtain robust statistical models that, as the name suggests, establish the relationship between the physicochemical properties of molecules and their biological activity to predict the behavior of new molecules. The fundamental principle is that the structural properties of molecules are reflected in their biological behavior.[10,11]

Thus, with current computational tools, it is possible to obtain drug candidates rationally, being faster and cheaper than experimental trials. The objective of this work is to develop reliable and convergent QSAR models for predicting novel cruzain inhibitors.[12]

2. MATERIALS AND METHODS

2.1. Data Set. The data set of cruzain inhibitors was obtained from articles within the literature. Compounds were chosen with the biological activities experimentally obtained by the same methodology to maintain uniformity through the data, later being organized in a table, having the structural information in simplified molecular input line entry system (SMILES) format, with their respective IC_{50} (which is the concentration of the compound that inhibits 50% of the biological activity), being converted to the logarithmic scale (pIC_{50}).

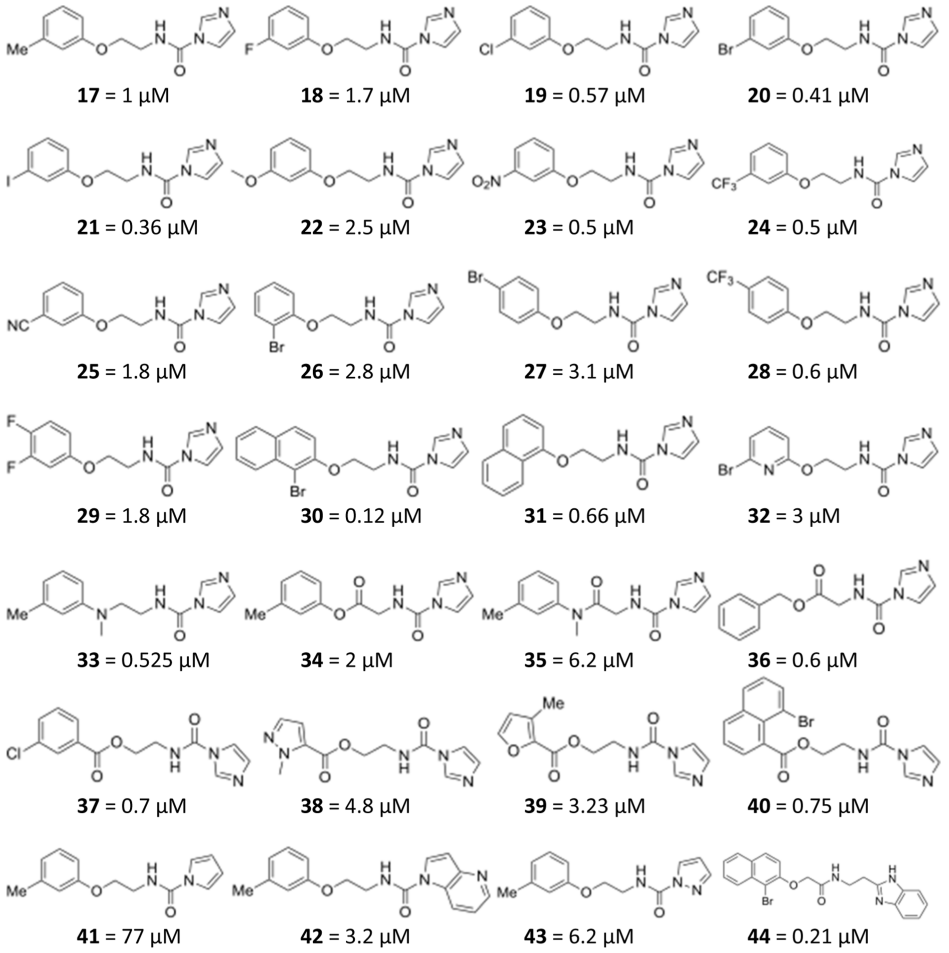
Three series of compounds from the same research group (Laboratory of Molecular Modeling and Drug Design) were selected. The first is composed of 33 molecules, mostly from cyclic imide class;[4] the second consists of 37 molecules derived from carbamoyl imidazoles;[13] and the third consists of 17 molecules derived primarily from triazine nitriles,[14] totaling 87 molecules with inhibitory activity for cruzain.

The activity of the compounds was measured by fluorescence spectroscopy under identical conditions. Of those, 61 compounds had IC_{50} data, which were used in the QSAR studies (Charts 1−3).

2.2. 2D-QSAR. After building the data set composed of cruzain inhibitors, molecular structures were converted from SMILES into 2D using the free tool "Online SMILES Translator" from computer-aided drug design (CADD) (Chemical Biology Laboratory (CBL), NCI, NIH, University of Erlangen-Nuremberg, Germany), available on https://cactus.nci.nih.gov/translate/.[15] Those molecules were submitted to the Padel-Descriptor software, which calculated 1444 one-dimen-

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Chart 2. Second Set Is Composed Mostly of Carbamoyl Imidazoles and Their Biological Activity (IC_{50})[13]



sional (1D) and 2D descriptors for each molecule in the data set.[16,17]

KNIME 4.7.1 (University of Konstanz, Zurich, Switzerland)[18] software was used for data organization, automatic/manual filtering of descriptors and molecules, generating seven different regression models, with internal validation by leave-one-out (LOO) and preliminary data analysis. Approximately, 80% of the data set was used as the training set (random selection, using "1995" as seed) and 20% for the test set (external validation). Y-scrambling was performed, with 100 variations for each model, generating values of YS (Y-scrambling) R^2 and YS Q^2. The detailed workflow and data sets can be found in the following GitHub Repository: https://github.com/Rafael-Bello-Goncalves/Convergent-QSAR-models-for-the-prediction-of-cruzain-inhibitors.

We conducted a preliminary virtual screening of a set of hydrazones to evaluate the potential of molecules that have not yet undergone experimental validation.

2.3. Docking. Molecular docking was performed to align the compounds in the active site of the protein. The model was validated by redocking the cocrystallized ligand. Structures of the compounds were converted from SMILES to 3D using the Online SMILES Translator.[15]

Gaussian 09 (Gaussian, Inc., Wallingford, CT, USA)[19] software was used for molecular optimization through the Semi-Empirical theory, with the PM6 method. Once the molecules achieved their lowest energy conformations, docking procedures were performed with GOLD software (GOLD, v2020 3.0, Cambridge Crystallographic Data Centre, Cambridge, UK).[30]

Crystals of cruzain 1ME4, with a resolution of 1.2 Å, bound to the inhibitor [1-(1-benzyl-3-hydroxy-2-oxo-propylarbamoyl)-2-phenyl-ethyl]-carbamic acid benzyl ester (T10)[20] and 3KKU, with a resolution of 1.28 Å, linked to the inhibitor *N*-[2-(1H-benzimidazole-2-yl)ethyl]-2-(2-bromophenoxy)acetamide (B95)[21] were used in studies of docking.

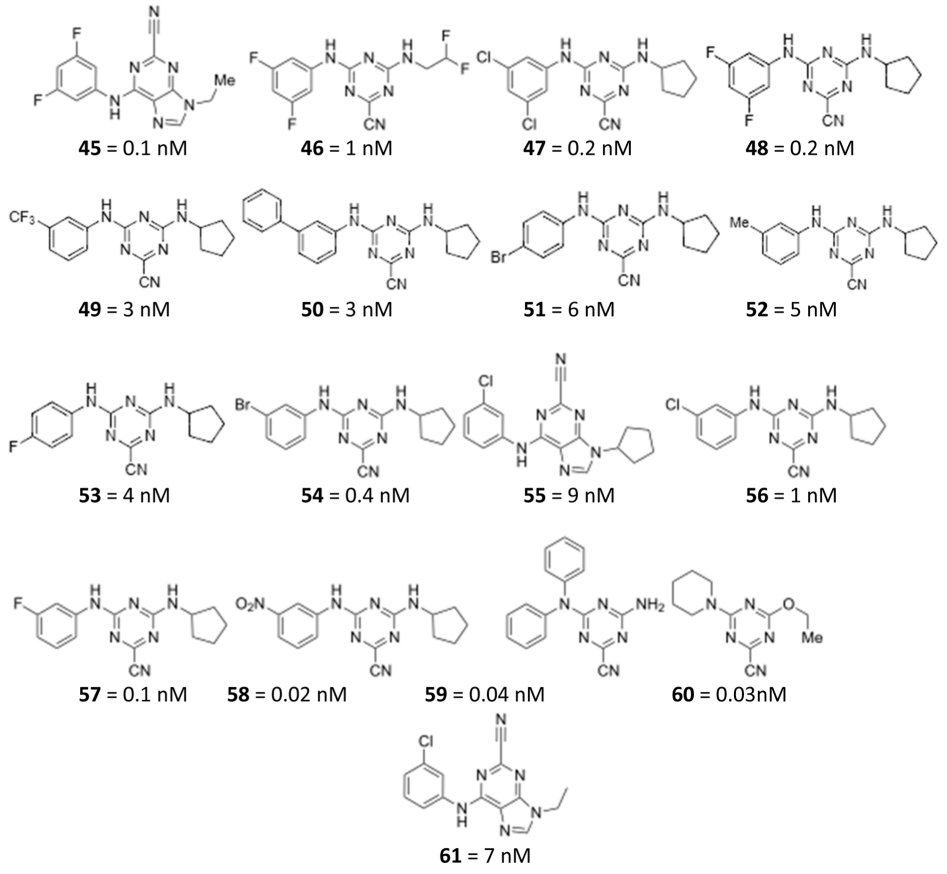
Both structures were prepared by removing ligands, water molecules, and other solvents and also removing the double conformations of amino acids. It was ensured that residue Cys_{25} was deprotonated, and His_{162} (in 3KKU) and His_{159} (in 1ME4) residues were fully protonated.

The ligand was treated as flexible, that is, with freedom for all rotatable connections, allowing a greater variety of poses and seeking the most adequate fit in the active site, while the enzyme was considered rigid. The active site was defined as the center of the sulfur atom of Cys_{25} for both structures, with a spacing of 10 Å in the three axes (*X*, *Y*, and *Z*) for the grid used in the calculations.

The ranking algorithm (score) used was ChemPLP, with ChemScore as rescore. Early termination was not allowed; therefore, the algorithm continues looking for better alternatives even after reaching an acceptable result. The selected search efficiency was 200%. Other configurations were kept at default values. These procedures were performed ten times for each compound, generating ten poses and choosing the best of three.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Chart 3. Third Set Is Composed Mostly of Triazine Nitriles Biological and Their Biological Activity (IC_{50}); Due to Their Elevated Potency, They Are Being Represented on the Nanomolar Scale, Instead of the Micromolar like the Previous Sets[14]



Each compound was evaluated using the top three solutions with the lowest RMSDs, and the alignment analysis was performed using UCSF ChimeraX 1.3 (Resource for Biocomputing, Visualization, and Informatics, University of California, San Francisco, USA)[22,23] software. The best conformation among the three possible solutions for each compound was manually selected based on the overlap with the original crystallized ligand structure.

2.4. 3D-QSAR. The 3D-QSAR studies were performed in SYBYL-X 2.1 (Certara Inc., Princeton, NJ, USA) and also in the python implementation of CoMFA (Py-CoMFA) available on the www.3d-qsar.com platform.[24−26]

For both, the preparation was the same: compounds already aligned by the docking process were loaded into the training and test sets (the same sets used in the 2D-QSAR study, so it is possible to compare different models). The formal charges were corrected, and experimental data were manually added to the platforms.

2.4.1. SYBYL (CoMFA and CoMSIA). The QSAR module of the SYBYL platform was used to select these data sets, starting with the training set, selecting the descriptor "comparative molecular field analysis (CoMFA)", considering both steric and electrostatic fields, with max/min energy cutoffs of 30 kcal/mol for each, with a smooth transition. Atomic charges were charged with Gasteiger, creating a model with pIC_{50} as the dependent variable and PLS analysis method in automatic mode, to choose the optimum number of principal components (up to six), and the validation of choice was LOO. The scaling was the CoMFA Standard. The probe atom was C.3, with a +1 charge. Thus, the values of CoMFA, Q^2_{LOO}, and R^2 were obtained. Then, the test set (external) was selected to perform the prediction using the previous model, obtaining the R^2 test value.

The same procedure was performed for CoMSIA, with the only difference being the descriptor (CoMSIA instead of CoMFA), considering the Steric, Electrostatic, Hydrophobic, Donor, and Acceptor Fields, with an attenuation factor of 0.3. Contour maps indicating the contributions of each field were generated for the best models of both CoMFA and CoMSIA. They were analyzed in SYBYL itself.

For both 3D models, we performed a y-scrambling stability test using the manual PLS option and selecting "Scrambling Stability Test"; as for the calculation parameters, we used the optimal number of principal components (after the automatic PLS results) of each model (CoMFA or CoMSIA). We performed 10 scramblings for each binning level with the maximum number of bins set to 10 and the minimum number of bins set to 2. The critical point was 0.85, and the seed was 1995. This generated 91 reports of YS R^2, YS $Q_{LOO}{}^2$, and YS SDEP that were analyzed in MS Excel.

2.4.2. Py-CoMFA. The Py-CoMFA application was used to build the CoMFA model. Similar to SYBYL, the Probe Atom was also C3.3, with a +1 charge. The dielectric Constant was eight, Molecular Interaction Field set to BOTH (evaluates both Steric and Electrostatic fields), a maximum number of principal components of eight, grid spacing of two, grid extension of five, max/min energy cutoffs 30 kcal/mol, minimum sigma of two, cross-validation in LOO, using compute unified device architecture (CUDA) to calculate the grid, and the option of applying the model to test set molecules set to True.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

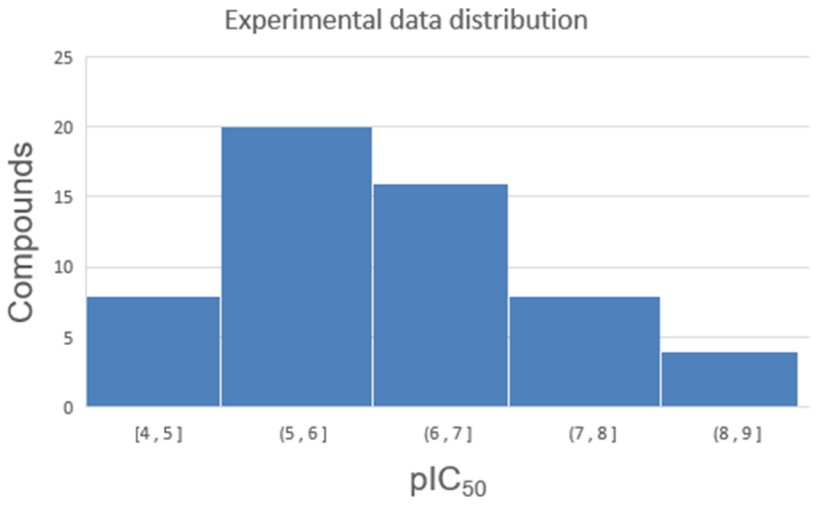


Figure 1. Histogram showing the distribution among the experimental data of the data set selected for the study.

Table 1. All Models Generated with the Full Data Set Using 61 Molecules[a]

model	descriptors	R^2	R^2 test	Q^2 LOO	s	YS R^2	YS Q^2	YS R^2 50%	YS Q^2 50%	YS R^2 99%	YS Q^2 99%
Random Forest a	1089	0.969	0.828	0.790	1.137	0.830	−0.179	0.831	−0.188	0.872	0.076
	112	**0.978**	**0.785**	**0.798**	**1.125**	**0.833**	**−0.153**	**0.832**	**−0.162**	**0.882**	**0.093**
	12	0.962	0.708	0.548	1.032	0.832	−0.160	0.829	−0.175	0.874	0.114
	5	0.948	0.782	0.365	1.110	0.827	−0.209	0.828	−0.213	0.869	0.122
	2	0.945	0.551	0.629	1.066	0.818	−0.250	0.820	−0.247	0.864	0.011
Tree Ensemble a	1089	0.965	0.853	0.789	1.148	0.842	−0.184	0.842	−0.195	0.881	0.143
	112	**0.979**	**0.817**	**0.801**	**1.121**	**0.846**	**−0.165**	**0.847**	**−0.166**	**0.879**	**0.152**
	12	0.966	0.710	0.581	1.025	0.843	−0.170	0.842	−0.184	0.884	0.080
	5	0.952	0.735	0.327	1.140	0.838	−0.218	0.838	−0.220	0.880	0.115
	2	0.951	0.600	0.615	1.082	0.829	−0.262	0.827	−0.270	0.873	0.056
Simple Regression a	1089	0.968	0.612	0.650	1.148	1.000	−1.167	1.000	−1.127	1.000	−0.281
	112	**1.000**	**0.690**	**0.764**	**1.221**	**1.000**	**−1.185**	**1.000**	**−1.177**	**1.000**	**−0.442**
	12	0.993	0.143	0.453	1.452	1.000	−1.089	1.000	−1.073	1.000	−0.216
	5	0.911	0.529	0.110	1.294	1.000	−1.168	1.000	−1.104	1.000	−0.231
	2	0.948	0.197	0.524	1.163	1.000	−1.015	1.000	−0.991	1.000	−0.180
Gradient-Boosted Trees a	1089	0.990	0.779	0.727	1.237	1.000	−0.517	1.000	−0.545	1.000	0.070
	112	**1.000**	**0.596**	**0.795**	**1.241**	**1.000**	**−0.504**	**1.000**	**−0.462**	**1.000**	**0.009**
	12	0.999	0.072	0.529	1.359	0.996	−0.467	0.999	−0.496	1.000	0.087
	5	0.981	0.554	0.219	1.213	0.997	−0.597	0.999	−0.609	1.000	−0.083
	2	0.948	0.447	0.584	1.138	0.990	−0.632	0.993	−0.602	0.999	−0.023
XGBoost a	**1089**	**0.915**	**0.523**	**0.406**	**1.317**	**1.000**	**−2.965**	**1.000**	**−2.766**	**1.000**	**−0.916**
	112	0.977	0.284	0.422	1.555	0.917	−3.401	0.926	−3.201	0.981	−0.746
	12	0.543	0.665	0.199	1.162	0.234	−0.458	0.226	−0.464	0.468	0.052
	5	0.309	0.392	0.183	1.096	0.066	−0.124	0.051	−0.136	0.221	0.061
	2	0.519	0.474	0.456	0.926	0.010	−0.049	0.003	−0.056	0.136	0.078
Linear Regression a	1089										
	112										
	12	0.542	0.647	0.175	1.165	0.253	−0.540	0.241	−0.551	0.491	0.009
	5	0.309	0.394	0.172	1.096	0.099	−0.184	0.084	−0.191	0.266	0.039
	2	**0.727**	**0.726**	**0.680**	**1.201**	**0.043**	**−0.090**	**0.025**	**−0.102**	**0.210**	**0.102**
Polynomial Regression a	1089										
	112										
	12	0.708	0.522	0.017	1.255	0.491	-1.6×10^{30}	0.479	-7×10^{29}	0.694	-1.5×10^{28}
	5	0.431	0.409	0.173	1.182	0.201	−1.043	0.183	−0.724	0.428	−0.069
	2	**0.689**	**0.685**	**0.604**	**1.166**	**0.083**	**−0.202**	**0.070**	**−0.174**	**0.271**	**0.072**

[a]The optimal number of descriptors is highlighted for each model. d = number of descriptors; R^2 = noncross-validated correlation coefficient; R^2_{test} = external test correlation coefficient; Q^2_{LOO} = leave-one-out cross-validation; s = standard deviation (external test set).

Variable pretreatment optimization was used to generate multiple models with different Charge Models, including Gasteiger, so it is possible to compare it to the SYBYL model and also analyze the best model generated by this platform,

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

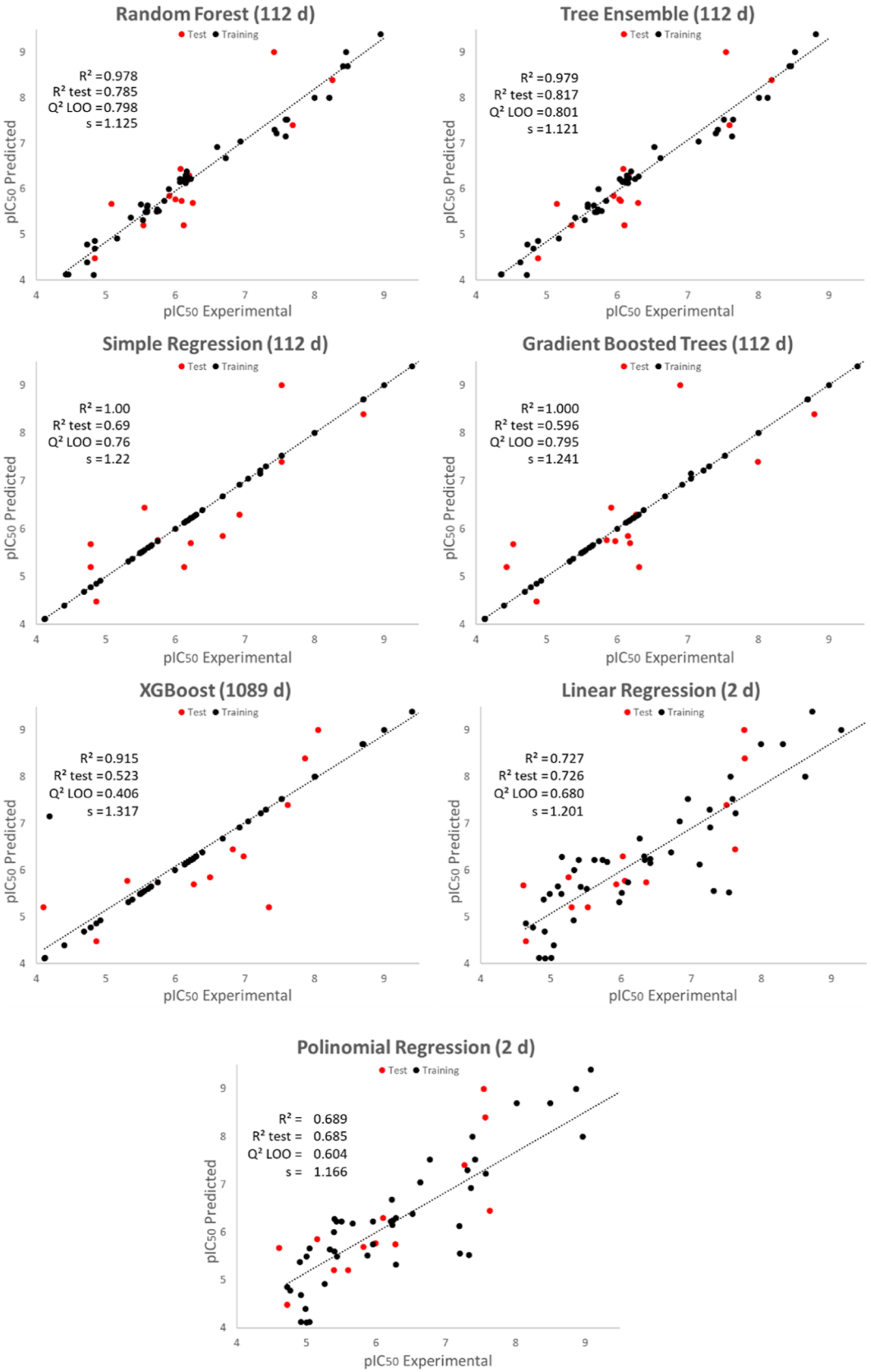


Figure 2. Best models from 2D-QSAR using the full data set (61 molecules). d = descriptors. R^2 = noncross-validated correlation coefficient; R^2_{test} = external test correlation coefficient; Q^2_{LOO} = leave-one-out cross-validation; s = standard deviation (external test set).

which was chosen considering the best Q^2_{LOO} from both steric and electrostatic fields. The contour maps were generated, and STDEV*COEFF maps were analyzed in ChimeraX for both Steric and Electrostatic Fields since they are more useful to see

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

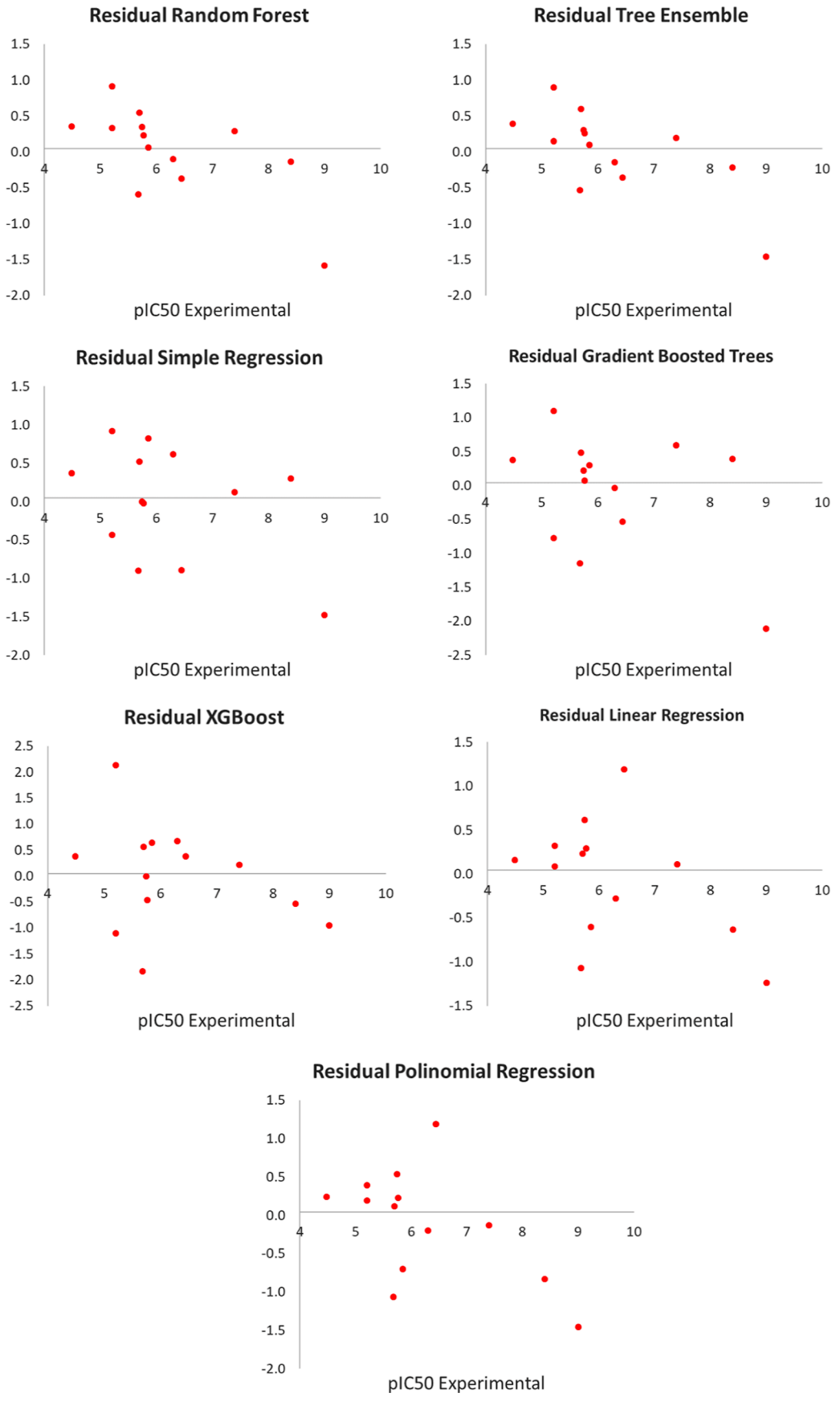


Figure 3. Residuals of the best algorithms for QSAR 2D (full data set).

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Table 2. Distribution between Training and Test Sets Used in All QSAR Studies, with Their Activity in the Logarithmic Scale (pIC_{50}), without Systematic Outliers [56 Molecules in Total; 44 in the Training Set (78,6%) and 12 in the Test Set (21,4%)][a]

Train.	pIC_{50}	Train.	pIC_{50}	Train.	pIC_{50}	Train.	pIC_{50}	test	pIC_{50}
1	5.38	**17**	6.00	**32**	5.52	**48**	8.70	**6**	5.85
2	5.64	**18**	5.77	**33**	6.28	**49**	7.52	**8**	5.66
3	4.92	**19**	6.24	**34**	5.70	**51**	7.22	**9**	4.86
4	5.60	**20**	6.39	**35**	5.21	**53**	7.40	**13**	5.68
5	6.22	**22**	5.60	**36**	6.22	**54**	8.40	**21**	6.44
7	6.22	**24**	6.30	**37**	6.15	**55**	7.05	**23**	6.30
10	4.40	**26**	5.55	**39**	5.49	**57**	9.00	**25**	5.74
11	4.78	**27**	5.51	**40**	6.12	**61**	7.15	**38**	5.32
12	4.69	**28**	6.22	**42**	5.49			**43**	5.21
14	4.12	**29**	5.74	**44**	6.68			**50**	7.52
15	4.12	**30**	6.92	**46**	8.00			**52**	7.30
16	4.48	**31**	6.18	**47**	8.70			**56**	8.00

[a]Train = training molecules; test = test molecules.

Table 3. 2D-QSAR Models, without Systematic Outliers[a]

model	descriptors	R^2	R^2 test	Q^2 LOO	s	YS R^2	YS Q^2	YS R^2 50%	YS Q^2 50%	YS R^2 99%	YS Q^2 99%
Random Forest b	1089	0.975	0.825	0.796	1.071	0.832	−0.196	0.831	−0.207	0.869	0.078
	112	0.974	0.867	0.772	0.968	0.834	−0.166	0.834	−0.178	0.872	0.114
	12	0.968	0.765	0.677	0.932	0.833	−0.172	0.833	−0.165	0.883	0.137
	5	**0.971**	**0.747**	**0.771**	**0.937**	**0.825**	**−0.224**	**0.823**	**−0.228**	**0.865**	**0.078**
	2	0.960	0.656	0.633	0.935	0.812	−0.282	0.810	−0.278	0.864	0.077
Tree Ensemble b	1089	0.972	0.813	0.805	1.097	0.847	−0.194	0.848	−0.203	0.879	0.093
	112	0.975	0.843	0.769	1.018	0.852	−0.164	0.852	−0.176	0.886	0.077
	12	0.967	0.836	0.653	0.938	0.849	−0.174	0.849	−0.171	0.886	0.117
	5	**0.973**	**0.789**	**0.793**	**0.955**	**0.842**	**−0.234**	**0.841**	**−0.252**	**0.882**	**0.090**
	2	0.956	0.669	0.655	0.953	0.827	−0.300	0.827	−0.295	0.872	0.045
Simple Regression b	1089	0.999	0.619	0.580	1.355	1.000	−1.139	1.000	−1.090	1.000	−0.168
	112	**1.000**	**0.748**	**0.624**	**1.354**	**1.000**	**−1.115**	**1.000**	**−1.126**	**1.000**	**−0.074**
	12	0.941	0.762	0.444	1.272	1.000	−1.113	1.000	−1.071	1.000	−0.384
	5	1.000	0.607	0.467	1.037	1.000	−1.165	1.000	−1.169	1.000	−0.289
	2	0.959	0.581	0.484	1.008	1.000	−1.066	1.000	−1.044	1.000	−0.298
Gradient-Boosted Trees b	1089	0.997	0.828	0.742	1.108	1.000	−0.546	1.000	−0.546	1.000	0.135
	112	0.982	0.816	0.751	1.162	1.000	−0.525	1.000	−0.536	1.000	0.348
	12	0.975	0.823	0.541	1.148	0.998	−0.516	0.999	−0.488	1.000	0.011
	5	**1.000**	**0.778**	**0.729**	**1.041**	**0.998**	**−0.588**	**0.999**	**−0.549**	**1.000**	**0.029**
	2	0.998	0.466	0.622	1.027	0.994	−0.685	0.996	−0.716	0.999	−0.146
XGBoost b	1089	0.982	0.707	0.552	1.421	1.000	−2.971	1.000	−2.610	1.000	−0.351
	112	**0.961**	**0.909**	**0.562**	**1.342**	**0.921**	**−3.446**	**0.931**	**−3.328**	**0.974**	**−0.502**
	12	0.783	0.692	0.381	0.983	0.247	−0.619	0.239	−0.543	0.498	−0.081
	5	0.373	0.543	0.137	0.779	0.080	−0.131	0.074	−0.125	0.245	0.066
	2	0.404	0.728	0.319	0.724	0.007	−0.054	−0.001	−0.063	0.089	0.030
Linear Regression b	1089										
	112										
	12	0.787	0.757	0.523	1.047	0.275	−0.806	0.261	−0.713	0.535	−0.057
	5	0.730	0.705	0.660	1.026	0.120	−0.204	0.116	−0.200	0.318	0.025
	2	**0.694**	**0.693**	**0.639**	**1.043**	**0.045**	**−0.104**	**0.031**	**−0.112**	**0.185**	**0.075**
Polynomial Regression b	1089										
	112										
	12	0.852	0.687	0.237	1.052	0.542	-2.3×10^{30}	0.549	-1.3×10^{30}	0.689	-5.6×10^{28}
	5	0.617	0.679	0.468	1.151	0.238	−1.191	0.236	−0.920	0.435	−0.242
	2	**0.620**	**0.637**	**0.543**	**1.108**	**0.093**	**−0.356**	**0.077**	**−0.237**	**0.265**	**−0.002**

[a]d = number of descriptors; R^2 = noncross-validated correlation coefficient; R^2_{test} = external test correlation coefficient; Q^2_{LOO} = leave-one-out cross-validation; s = standard deviation (external test set).

the differences between fields and target properties. Similar to the other models, y-scrambling was also performed, setting the "Make Y-scrambling" as true and the interactions to 10 (the maximum allowed in the platform).

2.4.3. HQSAR. The HQSAR models were built in SYBYL-X 2.1 using the HQSAR module and the same data set used in all studies of this work. PLS regression was also used to generate all

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

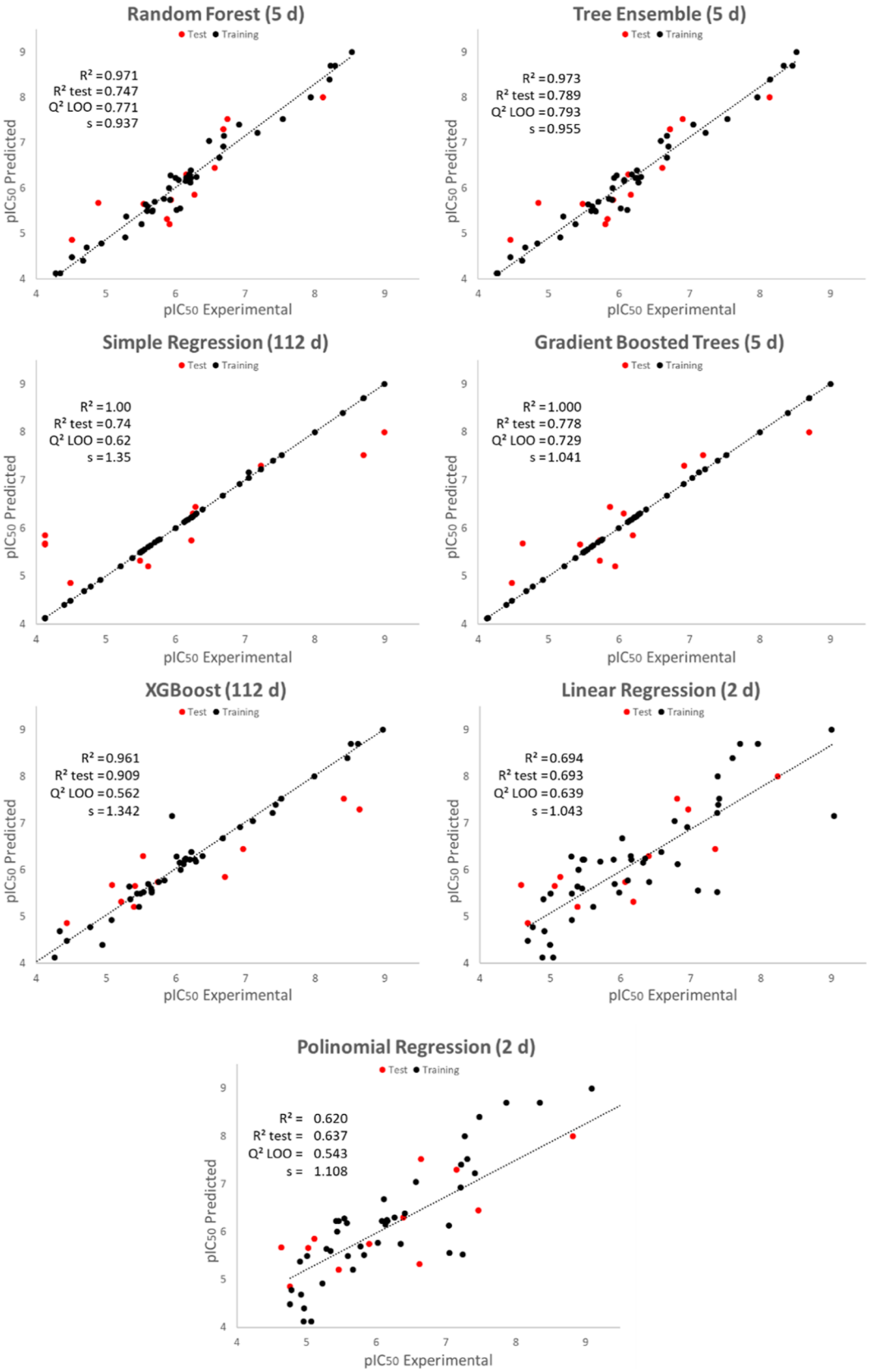


Figure 4. Best models from 2D-QSAR, without systematic outliers. d = descriptors. R^2 = noncross-validated correlation coefficient; R^2_{test} = external test correlation coefficient; Q^2_{LOO} = leave-one-out cross-validation; s = standard deviation (external test set).

of the models, with the optimum number of components determined by the LOO cross-validation procedure.

All of the default hologram lengths were used (53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, and 401 bins). Our first models

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

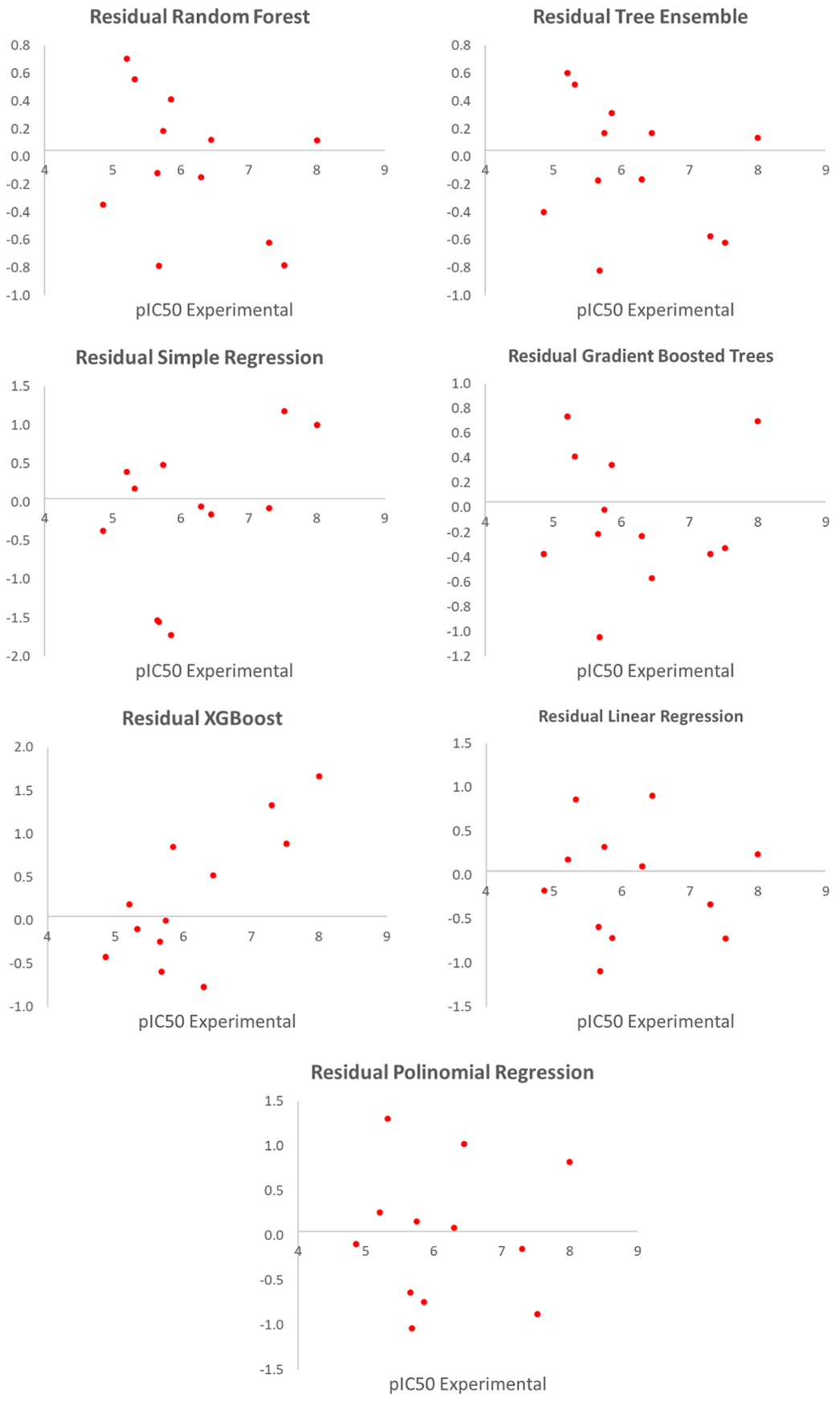


Figure 5. Residuals of the best algorithms for QSAR 2D (without outliers).

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

were built using the default four to seven atoms (4_7) count in fragments, using information from Atoms (A), Bonds (B), Connections (C), Hydrogen Atoms (H), Chirality (Ch), and Donor−Acceptor (DA) in different combinations. The best models were tested against different atom counts (2_5, 3_6, 5_8, and 6_9).

After the optimal settings were determined, a Y-scrambling test was conducted to assess the robustness of the HQSAR model. However, as SYBYL lacks a native option for scrambling HQSAR sets, the test had to be performed manually. Using KNIME, 100 random distributions of pIC50 values were generated. The HQSAR model was then applied to each distribution, and the resulting values (R^2, Q^2, and Ensemble of each run) were compiled manually into a table for further analysis.

3. RESULTS AND DISCUSSION

3.1. Data Set. One of the key aspects for a QSAR study to succeed is to have reliable experimental data.[27] In our study, we are using three different data sets[4,13,14] from the same research group. All compounds had their activity measured with the same methodology, and we chose to work with the IC_{50} values because they are statistically more robust (all compounds were measured at least in six different inhibitor concentrations, each one in triplicate). In addition, our data set has a logarithmic range of 5 orders of magnitude of the biological activity between the compounds (Figure 1).

Table 4. Contributions of Each Descriptor for the Linear Regression Model

set	AATS 5p (%)	AATS1i (%)
training	60	40
test	70	30

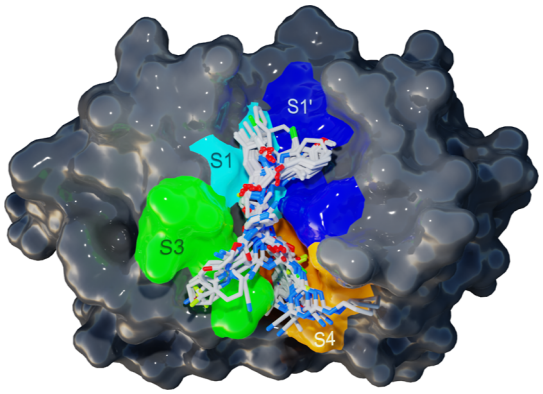


Figure 6. Alignment of all 61 inhibitors inside Cruzain (PDB: 1ME4) by GOLD. The colors within the cruzain are representations of its subsites. Cyan: S1 (Gln_{19}, Gly_{23}, Cys_{25}, Ser_{64}, Leu_{67}), dark blue: S1' (Asp_{158}, His_{159}, Trp_{177}), orange: S2 (Met_{68}, Ala_{133}, Leu_{157}, Gly_{160}, Glu_{205}), lime: S3 (Ser_{61}, Ser_{64}, Gly_{65}, Gly_{66}, Leu_{67}).[7,8]

3.2. 2D-QSAR. 2D-QSAR models were developed using KNIME, and the workflow is provided as Supporting Information, in addition to the SMILES structures.

We filtered the 1444 descriptors generated by Padel by removing all constant values, resulting in 1089 descriptors that were subsequently used to generate all of the 2D-QSAR models. Next, we added a correlation filter to remove descriptors that are highly correlated to each other, with different cutoffs, to determine the optimal number of descriptors for each model, the variations including 1089, 112, 12, 9, 5, and 3 descriptors. The initial data set was divided into 48 molecules in the training set (75.4%) and 13 molecules in the test set (24.6%), with random distribution.

LOO was chosen as the internal validation method to standardize this analysis across multiple platforms since other strategies like leave-many-out cannot guarantee the homogeneity of the data among multiple platforms, since not all support the use of seeds.

The best models overall are Random Forest and Tree Ensemble using 112 descriptors (Table 1, Figures 2 and 3). According to the literature, an acceptable QSAR model should have at least $R^2 = 0.600$, $R^2_{test} = 0.600$, and $Q^2_{LOO} = 0.500$, although higher values are desirable. When all of these values are met, Q^2_{LOO} usually has the highest priority. It should be noted that tendencies and patterns should also be avoided; therefore, predictions should have a random distribution.[28]

When we performed the Y-scrambling of the models, we found that the YS R^2 was the same as or lower than the R^2 of the models (except for XGBoost), and the YS Q^2 was always significantly lower (negative) than the original models. These results indicate that the models were not obtained by chance.

Although models such as Simple Regression and Gradient-Boosted Trees with 112 descriptors and XGBoost with 1089 descriptors meet the requirements of acceptable models (Table 1), they display signs of overfitting, which is when a model fits the training set very well (high value of R^2, usually with a perfect or near-perfect linear fit), but it performs poorly with unseen data (lower values of R^2_{test}).

One method of managing overfitting is to lower the number of descriptors. This seems to improve the pattern of the XGBoost model but lowers significantly either the R^2_{test} or Q^2_{LOO}, which are already low enough to discard this model, as it is not robust enough for predictions. Regarding the Simple Regression and Gradient-Boosted Trees, however, this perfectly linear pattern remains even when using as few as two descriptors (Supporting information). In this way, the two most recommended models for predictions are Random Forest and Tree Ensemble, with 112 descriptors.

Another observation is regarding the residual, which is the difference between the external set and the predicted activity. There is at least one molecule for each model that has a residual above one logarithmic unit, considered an outlier. The predicted activity values differed from the experimentally observed values

Table 5. Statistical Data of All of the QSAR 3D Models

model	R^2	R^2_{test}	Q^2_{LOO}	s	PC	YS R^2	YS Q^2	YS SDEP
Py-CoMFA a	0.975	0.794	0.747	1.023	5	0.897	0.062	1.145
Py-CoMFA b	0.996	0.786	0.810	0.987	8	0.959	−0.030	1.200
CoMFA (SYBYL)	0.939	0.728	0.680	1.004	3	0.887	0.568	0.901
CoMSIA (SYBYL)	0.961	0.833	0.799	1.065	4	0.906	0.517	0.835

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

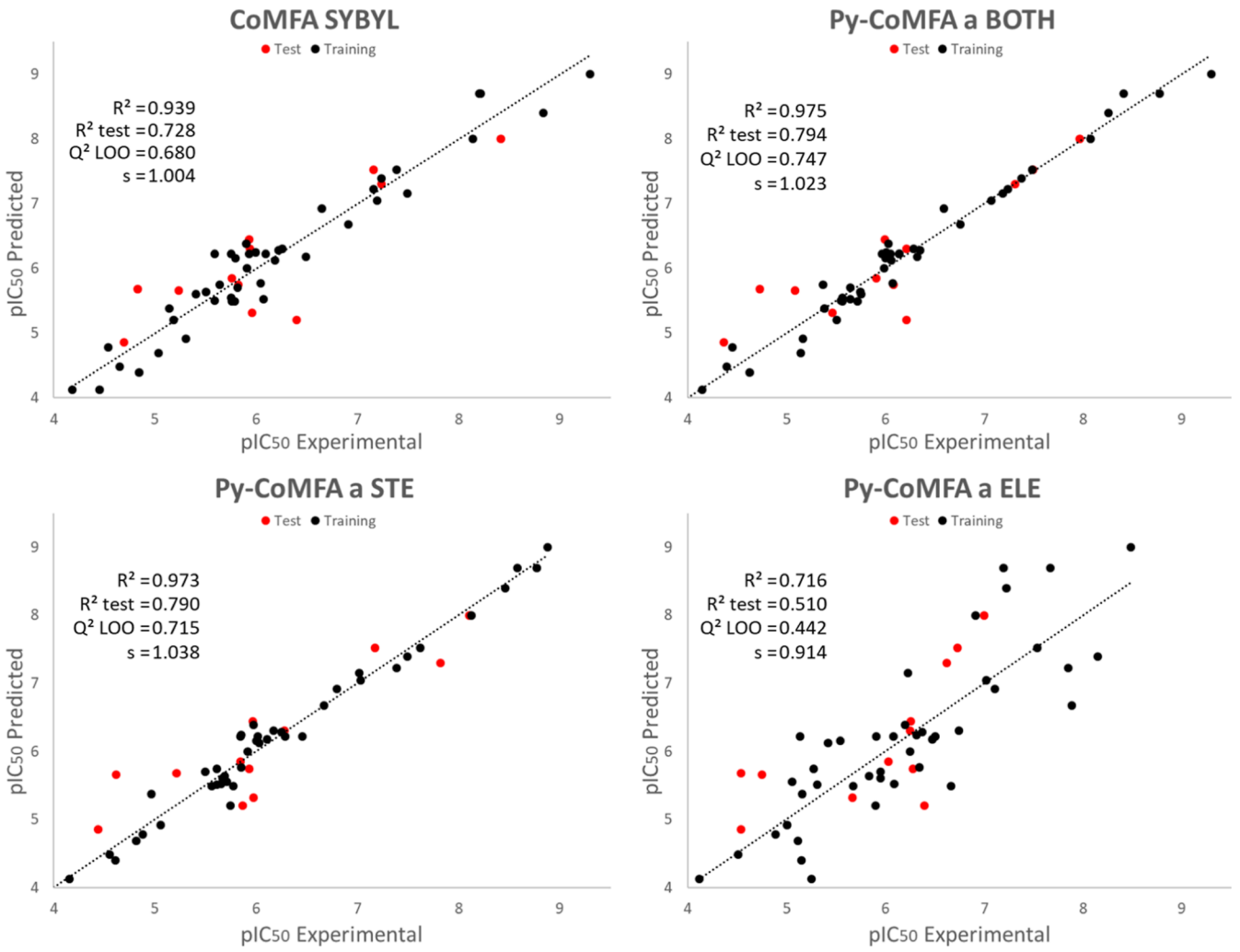


Figure 7. Comparison between CoMFA models from SYBYL and Py-CoMFA (Charge Model = gasteiger).

by more than one unit in the pIC_{50} in the regression models, that is, they did not contribute to the interpretation of the results.

After extensive testing with different models and descriptors (including other methodologies, like 3D-QSAR and HQSAR), we found five molecules with a systematic outlier behavior (for this study, we consider "systematic outliers" the molecules that perform very poorly regardless of the methodology or the distribution between training and test set, disrupting the prediction robustness of the model, with no positive contribution). As such, they were removed from further studies: **41**; **45**; **58**; **59**; and **60**.

Therefore, the filtered data set (without outliers) used for all the studies is listed in Table 2, also with the same seed for random distribution (1995), to avoid biases.

Without systematic outliers, a significant improvement can be observed through all models (Table 3, Figures 4 and 5).

Consistent with the previous Y-scrambling results, our analysis revealed that the YS R^2 values were equal to or lower than those of the models using the full data set. Additionally, the YS Q^2 values were significantly lower than those of the original models, suggesting that the models were not generated by chance.

The models with the best overall performance were Random Forest and Tree Ensemble, which exhibited a random distribution, had no apparent tendencies and did not show any residue above one logarithmic unit. It is worth noting that while using 1089 descriptors yields a higher Q^2_{LOO} for both, when compared with five descriptors [Q^2_{LOO} = 0.796 vs 0.771 for Random Forest, and Q^2_{LOO} = 0.805 vs 0.793 for Tree Ensemble (Figure 4)], the difference is marginal, and it is preferred to use the minimum descriptor, as it is easier to interpret five descriptors compared to 1089.

It should be noted that while Simple Regression with 112 descriptors and Gradient-Boosted Trees with five descriptors (Table 3) have decent statistical results, their training set still presents signs of overfitting (even though the R^2 test has performed well) and should be used with caution for the prediction of new cruzain inhibitors.

The XGBoost model has an acceptable performance with a tendency of higher deviations when the pIC_{50} increases and two outliers in the test set.

Linear Regression and Polynomial Regression exhibit a wide distribution in their training data (Figure 4), but with no tendencies, having a random distribution. They are optimal using two descriptors, being simple, but effective models to predict the activity of new cruzain inhibitors.

Most of the regression models can be explained with as few as five or two descriptors [except for Simple Regression and XGBoost, which have a better performance with 112 descriptors (Table 3)].

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

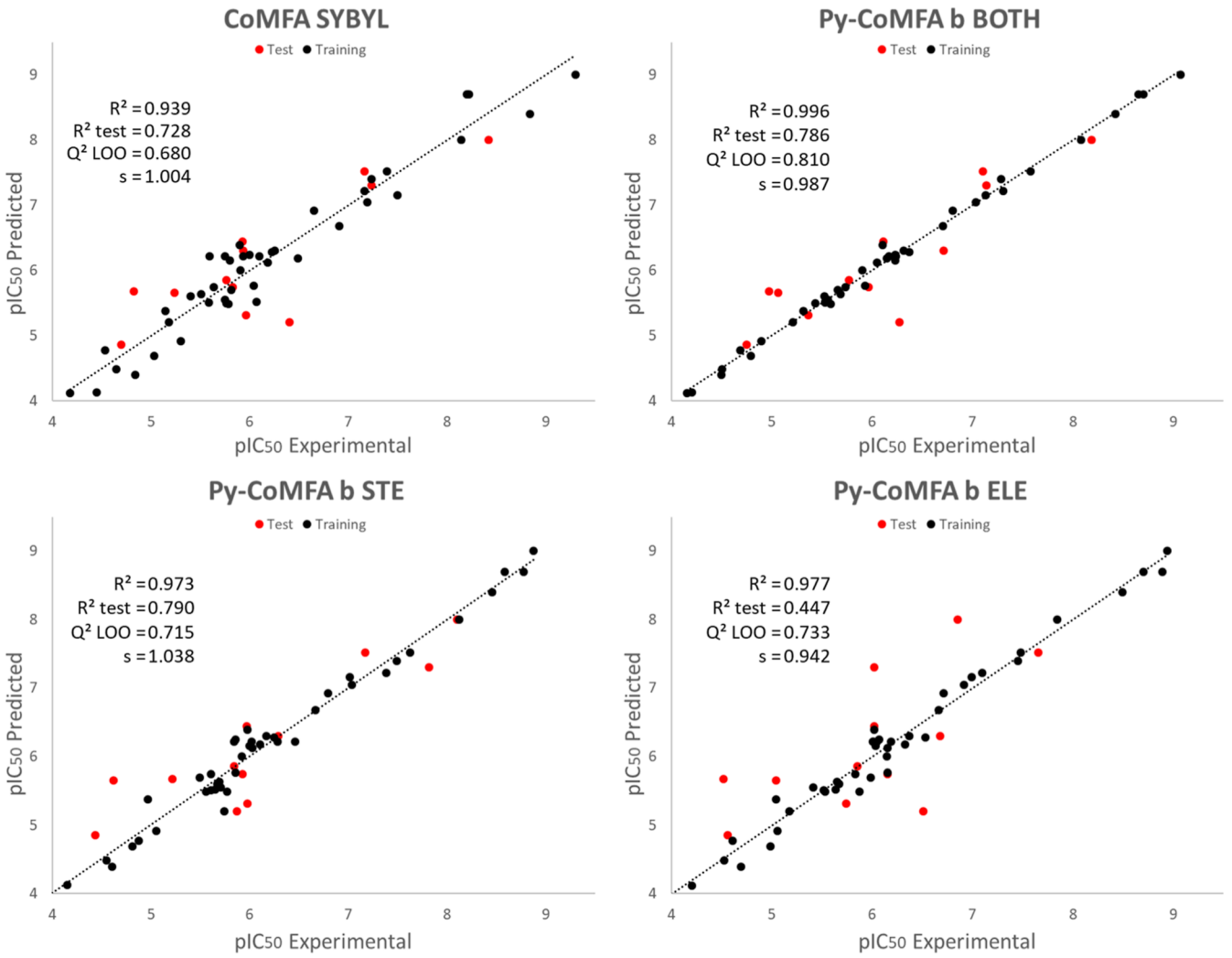


Figure 8. Comparison between CoMFA models from SYBYL and Py-CoMFA b (Charge Model = eem2015ha).

To study the relationship in our data set, we opted for Linear Regression a simpler model that successfully predicts the activity using only two descriptors, which are AATS 5p and AATS1i.

Using the R^2 value between each descriptor and the pIC_{50}, it is possible to obtain a linear relationship between both. Summing up all the descriptors R^2s and then taking all the fractions from each descriptor, we can obtain the contribution percentage of both (Table 4).

AATS 5p and AATS1i are autocorrelation descriptors. They belong to the class of topological descriptors, also known as molecular connectivity index, and describe the level of correlation between two objects (in this case, molecules) in terms of their specific structural property, or physicochemical property.[29]

AATS 5p (Table 4) was more relevant in both training and test sets and is an autocorrelation descriptor that calculates the average Broto-Moreau autocorrelation at lag 5, which is weighted by polarizabilities. It measures the average similarity between each atom in the molecule and its fifth nearest neighbor, taking into account their polarizability values. A higher AATS 5p value suggests that the molecule has a stronger autocorrelation pattern, which contributes to the biological activity. The higher the polarizability and first ionization potential, the higher the activity.[29]

AATS1i is an autocorrelation descriptor that calculates the average Broto-Moreau autocorrelation at lag 1, weighted by the first ionization potential (which is a measure of how easily an atom in a molecule loses an electron). A higher AATS1i value indicates that the molecule has higher activity and is more likely to lose an electron, potentially making it more reactive.[29]

3.3. Docking. Molecular alignment can be a very challenging step in 3D-QSAR studies. Fortunately, molecular docking is a technique that offers both accuracy and speed in finding the optimal conformations of ligands interacting with the receptor (cruzain).

Both inhibitors cocrystallized present on the selected cruzain structures (1ME4 and 3KKU) act noncovalently. This is important because the compounds used in QSAR studies also have the same mechanism of action, which is competitive.

The high resolution (1.2 Å) of the PDB 1ME4 crystal contributes to a greater precision in the conformation in the model used, and crystals with a resolution below 1.5 Å usually indicate a consequence of probably more than 95% of the observed data.[9]

It is important to use molecules in their minimal state of energy since molecular docking will generate multiple poses from this starting point. All compounds from the training and test series were optimized in Gaussian, using semiempirical (SE)

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

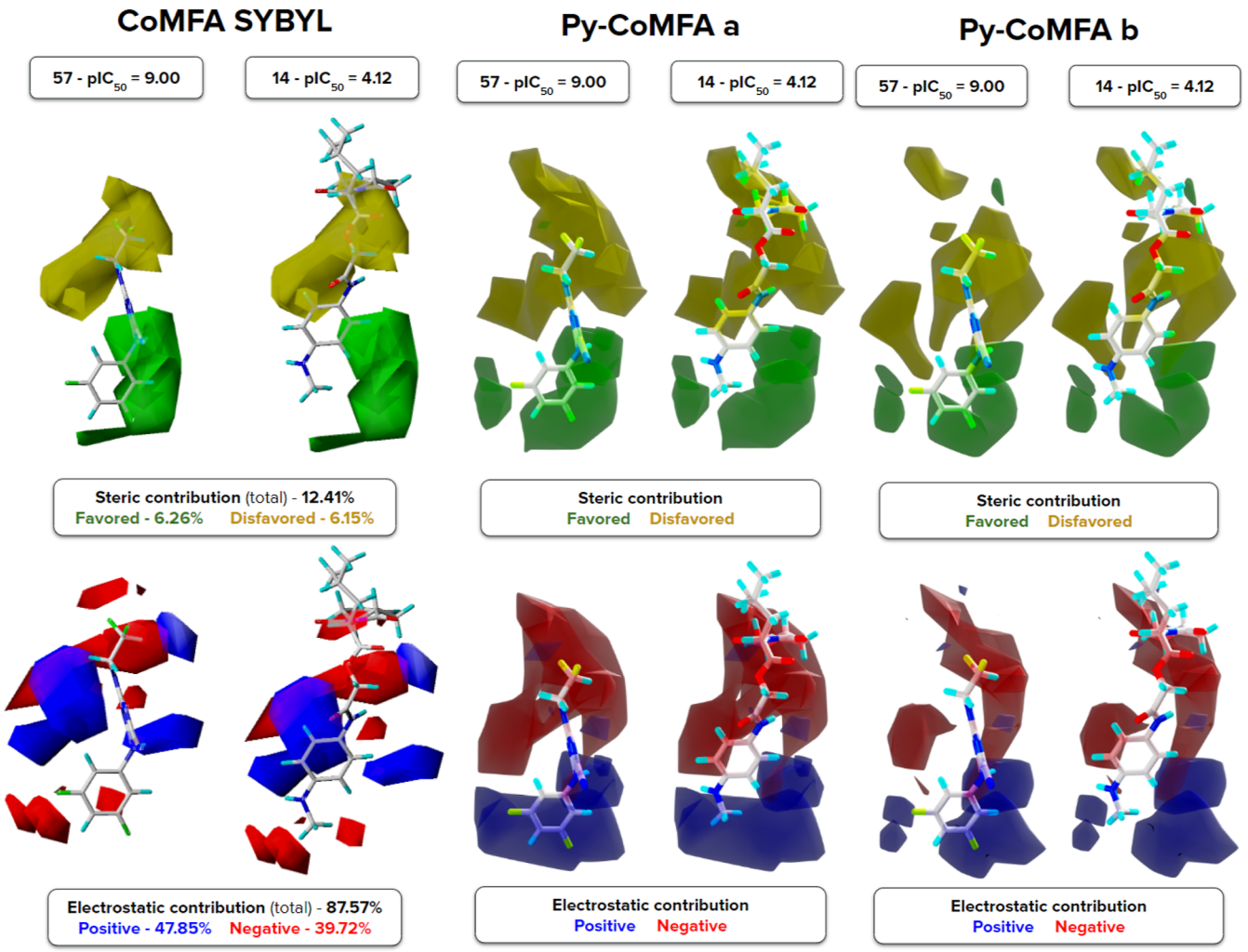


Figure 9. CoMFA contour maps from SYBYL, Py-CoMFA a (Gasteiger), and Py-CoMFA b (eem2015ha), highlighting steric and electrostatic effects, as well as their contributions to the model, when available.

theory, with the PM6 method. All of the structures converged correctly.

The best RMSD for redocking the cocrystallized ligand with the 3KKU structure was 43.0631 and was considered too high and therefore unsatisfactory to proceed with the alignment.

When redocking the cocrystallized ligand with the 1ME4 structure, the RMSD value was 2.4614. Considered to be satisfactory for correctly predicting the conformation obtained experimentally. Therefore, this methodology was used to align the ligands (Figure 6).

The alignment (Figure 6) shows good complementarity with the cruzain active site, and most of the residues were oriented between the Cys_{25} and His_{159} residues, which are part of the catalytic triad.

3.4. 3D-QSAR−CoMFA. After aligning all molecules in the data set, it was possible to perform both CoMFA and CoMSIA studies. This alignment is necessary because the models assume that all molecules are superimposed. The models compare the field of each molecule using PLS to identify and extract the contribution of chemical features of ligands with their biological activity.[11]

For the CoMFA model, two different platforms were used: SYBYL and Py-CoMFA (from www.3d-qsar.com). The first models shared the same parameters (with Gasteiger as the charge calculation method, to calculate atomic partial charges from the atomic coordinates of the molecules, denominated "a") to analyze whether the models would converge. Py-CoMFA also supports newer charge models, which were also tested to find the best model. We found that eem2015ha, a newer implementation derived from the electronegativity equalization method charge model to calculate atomic partial charges, denominated "b", was the best option when compared to Gasteiger, but it is not available in SYBYL for a direct comparison.

Since the Py-CoMFA does not provide the contribution percentage for steric and electrostatic fields (when considering both at the same time, in the same model), the statistical values from each field individually are also shown (Table 5, Figures 7 and 8) to check their robustness.

It can be observed that CoMFA from SYBYL and Py-CoMFA from a BOTH (considering both fields at the same time) have excellent statistical results (Figure 7), although the latter one is more robust, with less deviance from predicted activities. However, while the Y-scrambling results (Table 5) of all models have lower YS R^2s and YS Q^2s in comparison with the original ones, the CoMFA model generated by SYBYL did not demonstrate a substantial decrease in YS Q^2. As such, there is a possibility of chance correlation, and caution is advised when using it.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

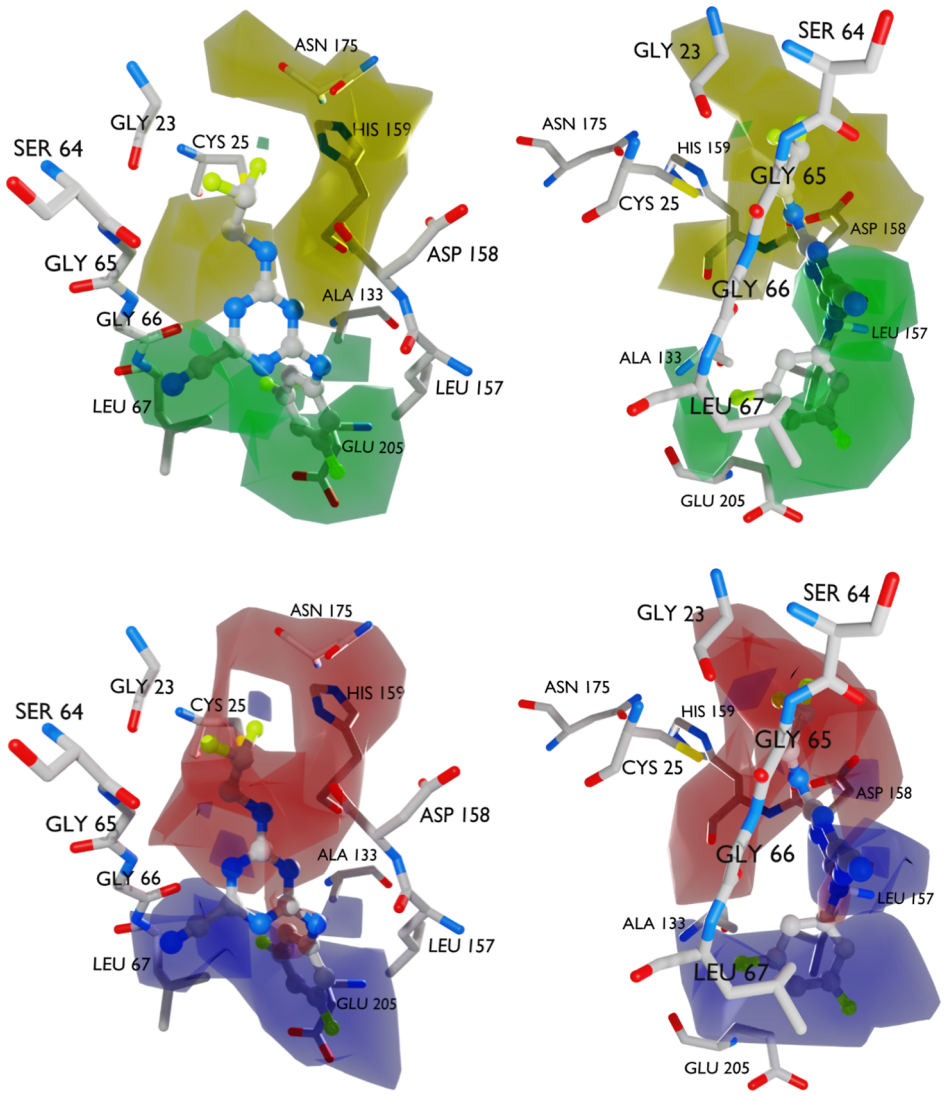


Figure 10. CoMFA contour models from Py-CoMFA b (eem2015ha) highlighting steric (above. yellow regions = disfavored | green regions = favored) and electrostatic (below. Blue regions = positive | red regions = negative) effects in the active site of cruzain.

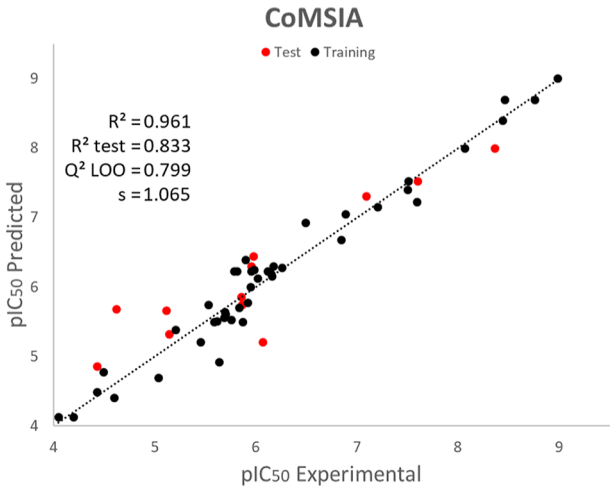


Figure 11. Statistical analysis of the CoMSIA model from SYBYL.

One interesting finding is that using only the Steric force field in Py-CoMFA can produce similar results compared to using both the Steric and Electrostatic force fields. In contrast, when using only the electric field, the statistical values are considerably worse (Figure 7).

Although Py-CoMFA does not output individual contributions from each field, considering the BOTH model, one can assume that the steric field contributes significantly more to the model than the electrostatic field.

Using the newer charge model (eem2015ha), an improvement can be observed while using both fields at the same time (Figure 8). Regarding only the steric field, the results are nearly identical to the model with Gasteiger (Figure 7), slightly changing only one test molecule but not enough to change the statistical results.

It can be noted that the individual electrostatic field had a noticeable improvement (Figure 8), although the predictions with the test set perform slightly worse (compared to the anterior with Gasteiger, Figure 7). This also suggests that the steric field has a greater contribution to the model than the electric field in this case.

3.5. 3D-QSAR−CoMFA Contour Maps. One advantage of 3D-QSAR is the possibility of contour map generation. This provides an easy way to visualize which regions from the model are favorable or unfavorable for biological activity.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

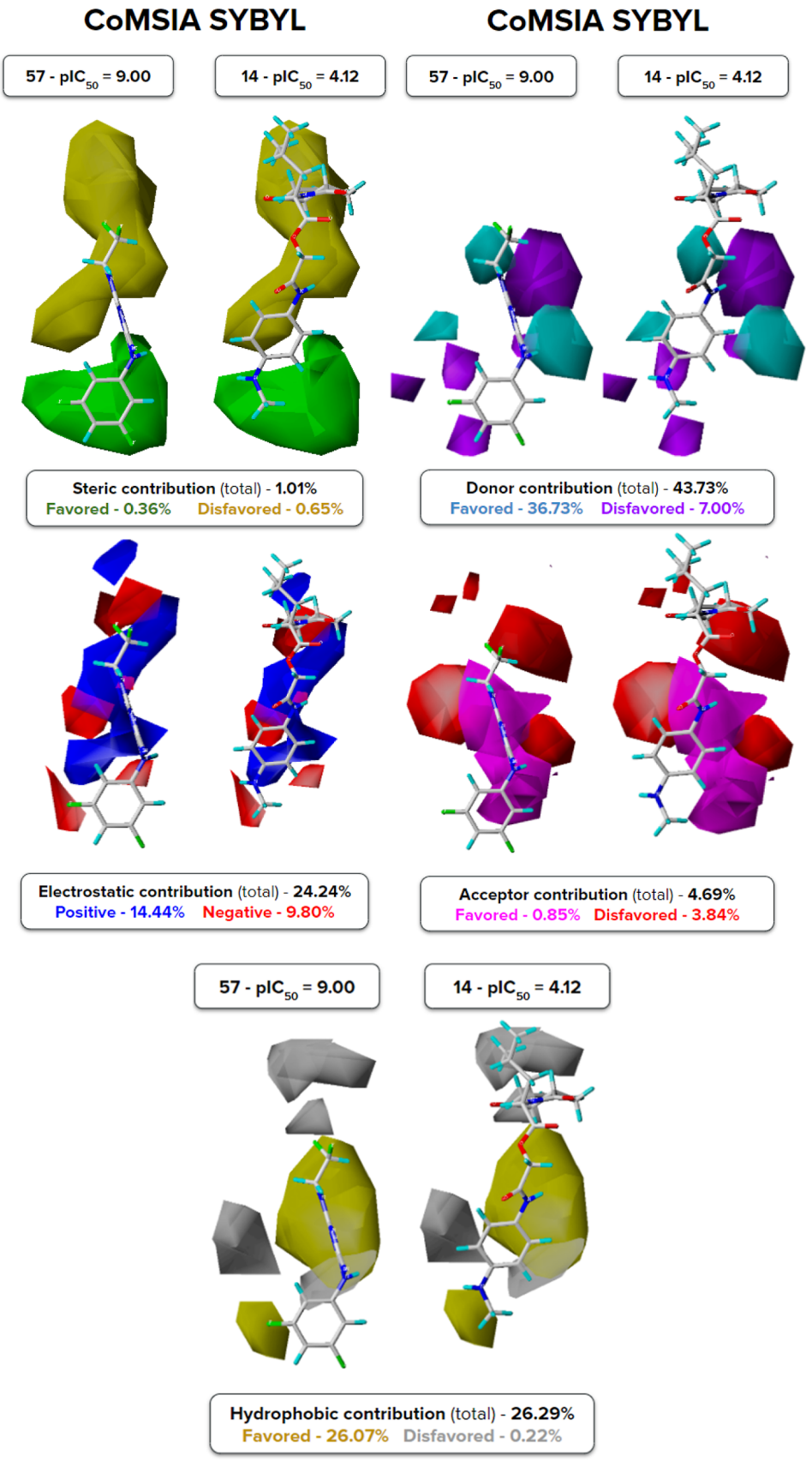


Figure 12. CoMSIA contribution maps from SYBYL, highlighting steric, donor, electrostatic, acceptor, and hydrophobic effects as well as their contributions to the model.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Table 6. HQSAR Results with Fragment Sizes between Four and Seven Atoms (4_7)[a]

model	Frag. distribution	Q^2_{LOO}	R^2	Ens	BL	PC
1	A/B	0.809	0.852	0.782	71	2
2	**A/B/C**	**0.824**	**0.879**	**0.794**	**59**	**3**
3	A/B/C/H	0.787	0.863	0.739	61	3
4	A/B/C/H/Ch	0.772	0.862	0.732	61	3
5	**A/C**	**0.824**	**0.868**	**0.799**	**97**	**2**
6	A/C/DA	0.806	0.861	0.776	401	2
7	A/C/H/Ch	0.764	0.888	0.742	353	4
8	**A/C/Ch**	**0.825**	**0.883**	**0.802**	**151**	**3**
9	A/C/Ch/DA	0.802	0.857	0.772	71	2
10	A/B/H	0.788	0.848	0.748	53	3
11	A/B/C/Ch	0.817	0.886	0.791	199	3
12	**A/B/C/DA**	**0.823**	**0.926**	**0.774**	**53**	**5**
13	A/B/H/DA	0.784	0.829	0.761	257	2
14	A/B/C/H/DA	0.765	0.910	0.736	307	4
15	A/B/H/DA	0.784	0.829	0.761	257	2
16	A/B/H/Ch/DA	0.763	0.882	0.736	257	3
17	A/B/C/H/Ch/DA	0.750	0.935	0.720	83	5

[a]Models n° 2,5,8 and 12 are highlighted. A = atoms, B = bonds, C = connections, H = hydrogen atoms, Ch = chirality, DA = donor−acceptor; Q^2LOO = leave-one-out cross-validation; R^2 = noncross-validated correlation coefficient. Ens = ensemble. BL = best length. PC = principal component.

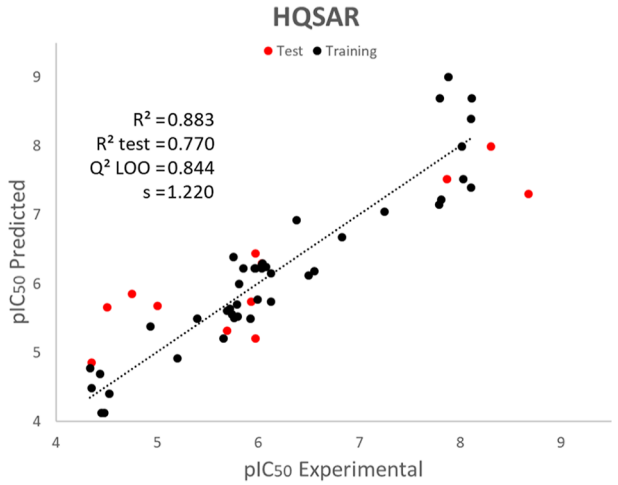


Figure 13. Statistical analysis of the HQSAR model from SYBYL.

All the models for CoMFA (Figure 9) were created considering both steric and electric fields at the same time because of their excellent statistical results discussed earlier. The most active molecule (**57**) is projected with the contours, as well as **14**, and although being the second least active component (**14** has IC_{50} = 75.0 μM, while **15** has IC_{50} = 75.5 μM), it has a more representative molecular alignment when compared to the majority of less active molecules, and only a marginal difference, compared to the **15** compound.

The favorable regions for steric interactions are colored in green, while unfavorable regions are colored in yellow (Figure 9). Positive charges which contribute to biological activity are shown in blue, while negative charges are shown in red. SYBYL also provides the percentage of contribution from each field that was also taken into account.

3.5.1. Steric Field. Steric contributions from all of the models converge with each other satisfactorily (Figure 9). It can be noted that the most active molecule (**57**) is oriented mostly in the favorable region, while the **14** molecule is oriented toward the unfavorable zones. An interesting fact is that the Steric field, when analyzed individually (not shown), is the same in Py-CoMFA "a" and "b"; however, the field changes slightly when considering BOTH, even when maintaining the same cutoffs for each field.

Table 7. HQSAR Results with Fragment Sizes Varying between Different Combinations of Atoms (Minimum Two Atoms and Maximum Nine Atoms)[a]

model 2 (A/B/C)	Q^2_{LOO}	R^2	Ens	BL	PC
2_5	0.818	0.867	0.790	353	2
3_6	0.816	0.863	0.793	401	2
4_7	0.824	0.879	0.794	59	3
5_8	0.824	0.893	0.791	97	3
6_9	0.820	0.901	0.776	97	3
model 5 (A/C)	Q^2_{LOO}	R^2	Ens	BL	PC
2_5	0.844	0.883	0.817	59	2
3_6	0.831	0.873	0.811	199	2
4_7	0.824	0.868	0.799	97	2
5_8	0.816	0.861	0.799	353	2
6_9	0.822	0.960	0.792	71	6
model 8 (A/C/Ch)	Q^2_{LOO}	R^2	Ens	BL	PC
2_5	0.846	0.884	0.818	59	2
3_6	0.826	0.873	0.811	59	3
4_7	0.825	0.883	0.802	151	3
5_8	0.822	0.865	0.792	353	2
6_9	0.810	0.848	0.786	353	2
model 12 (A/B/C/DA)	Q^2_{LOO}	R^2	Ens	BL	PC
2_5	0.813	0.887	0.788	151	3
3_6	0.815	0.859	0.778	61	2
4_7	0.823	0.926	0.774	53	5
5_8	0.798	0.916	0.767	151	4
6_9	0.809	0.923	0.760	83	4

[a]Model 5, with a fragment size of two to five atoms, was chosen as the best (highlighted). A = atoms, B = bonds, C = connections, H = hydrogen atoms, Ch = chirality, DA = donor−acceptor; Q^2_{LOO} = leave-one-out cross-validation; R^2 = noncross-validated correlation coefficient. Ens = ensemble. BL = best length. PC = principal component.

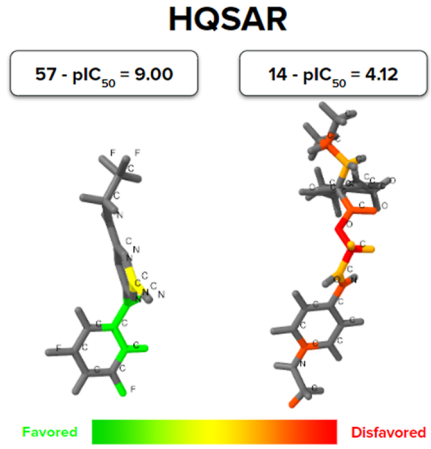


Figure 14. Contribution maps for the least active molecule (**14**) and the most active (**57**). Red represents disfavored areas, while green, the favored areas. Atoms are labeled, except for H atoms, for better clarity.

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Table 8. Best Models from This Work and the Statistical Analysis from Each One without Systematic Outliers[a]

model	d	R^2	R^2_{test}	Q^2_{LOO}	s	PC	n° out	n° out $_{test}$	YS R^2	YS Q^2
Random Forest	5	0.971	0.747	0.771	0.937		0	0	0.825	−0.224
Tree Ensemble	5	0.973	0.789	0.793	0.955		0	0	0.842	−0.234
Simple Regression	112	1.000	0.748	0.624	1.354		3	1	1.000	−1.115
Gradient-Boosted Trees	5	1.000	0.778	0.729	1.041		0	1	0.998	−0.588
XGBoost	112	0.961	0.909	0.562	1.342		2	1	0.921	−3.446
Linear Regression	2	0.694	0.693	0.639	1.043		5	1	0.045	−0.104
Polynomial Regression	2	0.620	0.637	0.543	1.108		3	4	0.093	−0.356
CoMFA (SYBYL)		0.939	0.728	0.680	1.004	3	0	1	0.887	0.568
Py-CoMFA a		0.975	0.794	0.747	1.023	5	0	1	0.897	0.062
Py-CoMFA b		0.996	0.786	0.810	0.987	8	0	1	0.959	−0.030
CoMSIA (SYBYL)		0.961	0.833	0.799	1.065	4	0	1	0.906	0.517
HQSAR (SYBYL)		0.883	0.770	0.844	1.220	2	1	3	0.372	0.053

[a] d = number of descriptors; R^2 = noncross-validated correlation coefficient; R^2_{test} = external test correlation coefficient; Q^2_{LOO} = leave-one-out cross-validation; s = standard deviation (external test set); PC = principal component; n° out = number of outliers from training set; n° out $_{test}$ = number of outliers from test set. YS R^2 = noncross-validated correlation coefficient of the y-scrambling. YS Q^2 = leave-one-out cross-validation of the y-scrambling.

Even though the Steric field seems to have the biggest impact in Py-CoMFA models, for SYBYL, the Steric contribution is much lower, with 12.41% in total.

Observing Py-CoMFA b (both) and the **57** molecule superimposed in the active site (Figure 10), most of the unfavorable zone is found within the looping area inside the cruzain, which has little interaction with the solvent, while the favorable zone is free for more voluminous substituents to be inserted. This may help not only with biological activity but also with selectivity since fewer enzymes would have this free area available.

More specifically, the bulk of the favored area can be found between Leu_{67}, Met_{68}, Leu_{157}, and Glu_{205}, being the last tree residues especially important, as they are part of the S2 subsite, the most relevant subsite for the enzyme specificity. Regarding the disfavored area, it is found mostly around the Gly_{23}, Cys_{25}, Ser_{64}, Gly_{65}, Gly_{66}, Ala_{133}, Asp_{158}, and His_{159} residues which covers a great portion of S1 and S1′ subsites. It should be mentioned that Cys_{25} and His_{159} are two residues from the catalytic triad. Similar findings were also mentioned, in earlier studies, with a different data set.[9]

3.5.2. Electrostatic Field. The electrostatic field from SYBYL does not match Py-CoMFA (Figure 9) completely, even with the same Charge Model (Gasteiger, in Py-CoMFA a model). This may be explained by differences in the algorithm found in each platform. It also may help explain why for SYBYL the electrostatic contribution is so important when the statistical data in Py-CoMFA suggest otherwise.

However, there are two regions where they intersect; the negative (in red) comprehends the looping region inside the enzyme, especially near the S1 and S1′ subsites, while the positive (in blue) region is near the Gly_{66} residue of the S3 subsite.

The main differences are that for SYBYL, positive charges are preferred closer to Cys_{25}, and negative charges are preferred for Leu_{67}, Leu_{157}, and Glu_{205}; however, for Py-CoMFA, the opposite is true for these residues.

3.6. 3D-QSAR−CoMSIA. At the moment of writing this article, there is not a CoMSIA methodology available on www.3d-qsar.com; therefore, we performed all analyses within SYBYL. The statistical values can be found in Figure 11.

The CoMSIA model presents satisfactory statistical data (Table 5 and Figure 11), and it is appropriate for further analysis.

3.6.1. 3D-QSAR−CoMSIA Contour Maps. Following the same methodology from the CoMFA contour maps, we can observe the Steric, Donor, Electrostatic, Acceptor, and Hydrophobic contributions (Figure 12).

Zones from the Steric field are very similar to those found previously in CoMFA studies, although the contribution to the model is minimal, with 1.01% in total.

Donor contributions, in contrast, are the most relevant for CoMSIA, especially with favored interactions (36.73%, in light blue), highlighting Gly_{66}, which is near the nitrile from the **57** molecule, as well as the benzene ring from **14**, and Gly_{23}, near the halogens from **57**, as well as the ester portion of **14**.

The electrostatic field, once again, does not match the others found in CoMFA studies. Therefore, caution is advised when planning new inhibitors using this type of field as a reference. It contributes 24.34% of total relevance to the model, and although it switches between negative and positive zones, the positive tends to be found inside the ligands and the negative for the areas closer to the residues in general.

Acceptor contributions are marginal (4.69% in total), favoring the regions toward the S2 subsite, and disfavoring the regions toward the S1 and S1′ subsites.

Hydrophobic interactions are mostly favored (26.07%, in yellow), which is closer to Leu_{67} (S3 subsite, secondary amine on **14**), as well as S1 and S1′ subsites.

3.7. HQSAR. The first HQSAR models were made using all available and default hologram lengths provided by SYBYL (53-401 bins), with fragment sizes from four to seven atoms (4_7). Our objective was to analyze which fragment distribution was the optimal choice considering the Q^2_{LOO}, R^2, and Ensemble values for each variant (Table 6).

All of the models (including the least performing ones) have excellent statistical data (Table 6). Models n° 2,5,8 and 12 were chosen to continue the study with different fragment sizes (Table 7).

By conducting a Y-scrambling test on the optimal model (A/C and 2_5 fragments), the following results were obtained: YS R^2 = 0.372, YS Q^2 = 0.053, and YS Ensemble = 0.034. These values demonstrate that the model was not obtained by chance since they differ significantly from the original results (Table 7).

Although model 8 (A/C/Ch), with a fragment size between two and five atoms (Table 7), has a marginal improvement over model 5 (A/C) with the same fragment size, it is preferred to use

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

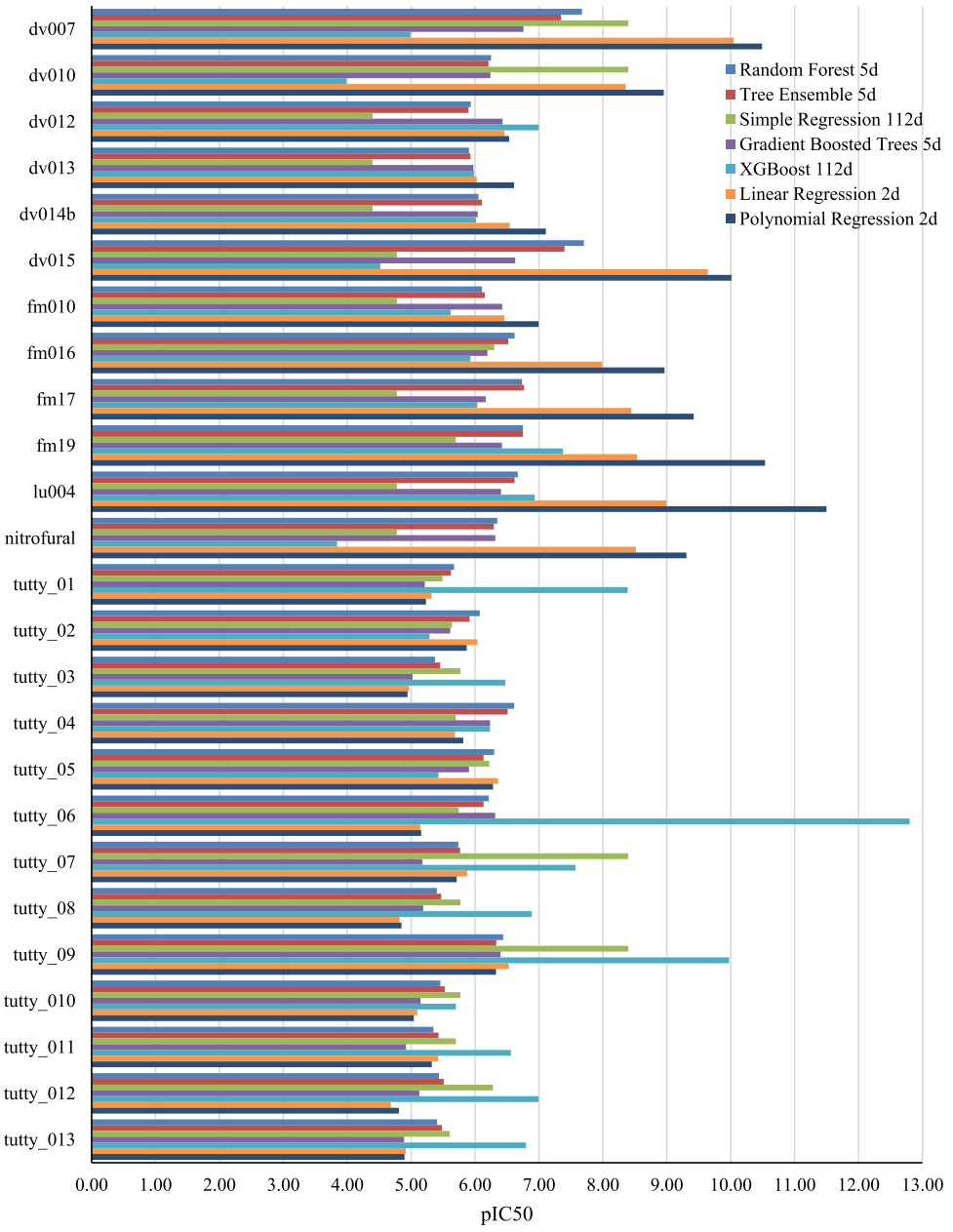


Figure 15. Predicted pIC50 values for the Virtual Screening of hydrazones.

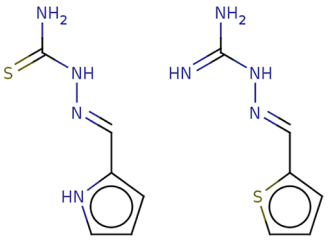


Figure 16. Most promising hydrazone molecules selected from the Virtual Screening: dv007 (left) and dv015 (right).

model 5 for HQSAR, since it requires only information about Atoms (A) and Connections (C), with no need for Chirality (Ch). The analysis of this model is shown in Figures 13 and 14.

The HQSAR model brings decent results (Figure 13); however, there is also high deviance (s = 1.220) and tree outliers in the test set.

Observing the contribution maps (Figure 14), the most important in **57** belongs to the benzene ring, with halogens (F), which is located between Leu_{67}, Met_{68}, Leu_{157}, and Glu_{205} residues corroborating with the favorable zone from the steric field found in previous studies with CoMFA and CoMSIA contour maps (Figures 9, 10, and 12).

The nitrile from the **57** compound contributes to the biological activity, which is in proximity with Gly_{66}, and also corroborates previous findings in CoMFA and CoMSIA contour maps, fulfilling especially steric and donor favorable fields. Interestingly, though most electrostatic fields disagree with each other, this is the only region that has convergence from all of them in the positive region (Figures 9, 10, 12, and 14).

Ester and carboxamide groups, as well as the isobutane, stand out in the middle of the **14** molecule as a disfavored area, near Gly_{23}, Cys_{25}, Ser_{64}, Gly_{65}, Asp_{158}, and His_{159} residues, which once

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

Table 9. Predicted pIC_{50} Values for the Virtual Screening of Hydrazones[a]

mol	Random Forest 5d	Tree Ensemble 5d	Simple Regression 112d	Gradient-Boosted Trees 5d	XGBoost 112d	Linear Regression 2d	Polynomial Regression 2d	Average 3 best (RF, TE, and GBT)
dv007	**7.67**	**7.34**	**8.40**	**6.76**	**4.99**	**10.05**	**10.49**	**7.26**
dv010	6.25	6.21	8.40	6.24	3.99	8.36	8.95	6.23
dv012	5.93	5.90	4.40	6.43	7.00	6.46	6.53	6.09
dv013	5.90	5.93	4.40	5.97	5.98	6.03	6.61	5.94
dv014b	6.05	6.11	4.40	6.04	6.01	6.54	7.11	6.07
dv015	**7.70**	**7.40**	**4.78**	**6.63**	**4.52**	**9.64**	**10.01**	**7.24**
fm010	6.11	6.15	4.78	6.42	5.62	6.46	6.99	6.23
fm016	6.62	6.52	6.30	6.19	5.93	7.99	8.97	6.44
fm17	6.73	6.77	4.78	6.17	6.04	8.44	9.42	6.56
fm19	6.75	6.75	5.70	6.42	7.38	8.53	10.54	6.64
lu004	6.67	6.62	4.78	6.41	6.93	8.99	11.50	6.56
nitrofural	6.35	6.30	4.78	6.32	3.84	8.52	9.31	6.32
tutty_01	5.67	5.62	5.49	5.21	8.39	5.32	5.23	5.50
tutty_02	6.07	5.91	5.64	5.61	5.29	6.04	5.87	5.87
tutty_03	5.37	5.45	5.77	5.02	6.48	4.97	4.94	5.28
tutty_04	6.61	6.51	5.70	6.23	6.23	5.69	5.82	6.45
tutty_05	6.30	6.13	6.22	5.91	5.43	6.36	6.28	6.11
tutty_06	6.21	6.13	5.74	6.31	12.80	5.14	5.16	6.22
tutty_07	5.74	5.77	8.40	5.18	7.57	5.87	5.71	5.56
tutty_08	5.40	5.47	5.77	5.19	6.88	4.82	4.85	5.35
tutty_09	6.44	6.33	8.40	6.40	9.97	6.53	6.33	6.39
tutty_010	5.46	5.52	5.77	5.15	5.70	5.10	5.04	5.38
tutty_011	5.35	5.43	5.70	4.92	6.56	5.42	5.32	5.23
tutty_012	5.43	5.51	6.28	5.13	6.99	4.68	4.81	5.36
tutty_013	5.41	5.48	5.60	4.89	6.80	4.92	4.90	5.26

[a] d = number of descriptors; RF = random forest; TE = tree ensemble; GBT = gradient-boosted trees.

again corroborates with all the disfavored steric fields (Figures 9, 10, 12, and 14).

The carbon next to the secondary amine in the **14** compound is not in good agreement with some of the previous contour maps, as it is located near Gly_{66}, in an area favored by steric forces, as well as (in most cases) positive areas in the electrostatic field. However, it agrees with the hydrophobic forces for having a polar charge and is unfavorable, and with the donor map, as by being an electron donor group, it is located in the unfavorable area (Figures 9, 10, 12, and 14).

3.8. Comparison of All Models. Summing up all the model variations produced in this work, there were a total of 115 model variations. Out of those, 70 models were from 2D-QSAR (35 using the full data set, 35 without the outliers), 37 models from HQSAR (17 varying the distribution of the fragments, 20 varying the fragment sizes), and 8 models from 3D-QSAR (7 from CoMFA (including the isolated steric and electrostatic fields), and 1 from CoMSIA).

Out of all the 115 variations, there are 76 that match the minimum requirements to be considered as acceptable (R^2 = 0.600; R^2_{test} = 0.600; Q^2_{LOO} = 0.500). Of those, 33 are from 2D-QSAR (11 in the full data set and 22 without the outliers), 37 from HQSAR, and 6 from 3D-QSAR.

The best variations from each technique are shown in Table 8, considering the data set without systematic outliers.

All models (Table 8) met the minimum requirements to be acceptable QSAR models, with little to no outliers (which are considered to be residuals with a difference of one logarithmic unit) in both training (44 molecules) and test (12 molecules) sets. However, due to signs of overfitting, Simple Regression and Gradient-Boosted Trees should be used with caution.

Regarding the Virtual Screening, we used a small set of hydrazones for their similarity with our data set in our QSAR 2D study. We leveraged the combined predictive power of our top three 2D models: Random Forest, Tree Ensemble, and Gradient-Boosted Trees, all utilizing 5 descriptors each, as they were some of the best models of this work. This method helped identify the most promising molecules for future experimental validation.

Among the candidates (Table 9 and Figures 15 and 16), dv007 and dv015 consistently exhibited the highest predicted pIC50 values across our top-performing models. These molecules, with average pIC_{50} scores of 7.26 and 7.24, respectively, stand out as the most compelling candidates for experimental validation, suggesting their potential for further exploration in drug development.

Finally, it is essential for us to establish a correlation between our work and other recent publications on QSAR involving cruzain. In this context, we would like to highlight the research conducted by Rosas-Jimenez et al.,[31] where they employed various sets of cruzain inhibitors from diverse sources, resulting in a final data set comprising 344 compounds. They utilized the k-nearest neighbors and random forest algorithms to create both local and global models. The statistical parameters for the internal and external validation of their models indicated a significant level of predictability. Additionally, they defined the applicability domain quantitatively using leverage and similarity methods.

While our data set contains fewer structures compared to theirs, leading to a smaller molecular diversity, all our compounds originate from three series of tests conducted under the same protocol and by the same research group. This reinforces the method's robustness and adherence to good

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982

QSAR practices. In addition to the Random Forest, we also incorporated various QSAR techniques, including HQSAR, CoMFA, and CoMSIA, which exhibit great alignment among them. As a result, we believe that our work can be considered a robust reference for building quantitative models based on cruzain inhibitors, complementing the aforementioned study.

4. CONCLUSIONS

A comprehensive QSAR study was carried out in this work to create robust, predictive models for cruzain inhibitors. It is crucial to note that our work is entirely theoretical. The compounds used as the foundation for these models were experimentally tested by the Laboratory of Molecular Modeling and Drug Design, ensuring data uniformity.

The 2D-QSAR models are effective even with just two or five descriptors, making them useful for understanding which alterations can improve biological activity and save computational resources. Utilizing 2D-QSAR models in KNIME provides a free, open-source platform that is accessible to everyone for testing new ligands. The Random Forest, Tree Ensemble, and Linear Regression models demonstrated excellent predictive ability and high levels of precision, despite using a relatively small number of descriptors. However, caution may be necessary for the Gradient-Boosted Trees model due to potential overfitting.

The HQSAR models combine 2D-QSAR speed with some benefits of 3D-QSAR, including fragment contour maps for better and worse contributions. However, due to its high standard deviation (s = 1.220), it may lead to less accurate predictions.

Our 3D-QSAR models had great statistical results, with excellent convergence in all Steric Fields. However, the Electrostatic Fields had conflicting results, except for the Gly66 residue, which benefits from positive substituents. Caution is advised when using the CoMFA model from SYBYL due to minimal differences between y-scrambling results and the original model. The CoMSIA model identifies hydrogen donor substituents as the primary contributors to biological activity at 36.73%, followed by hydrophobic substituents at 26.07%. Despite their predictive power, generating and optimizing 3D structures and alignments require time and optimization and should be considered when developing new antichagasic agents.

In our preliminary virtual screening with hydrazones, based on our top three 2D models (Random Forest, Tree Ensemble, and Gradient-Boosted Trees, each with 5 descriptors), dv007 and dv015 emerge as promising candidates for further investigation. However, it is essential to note that these findings have not yet been experimentally validated.

These models should contribute to the community both as filters for virtual screening and as a way to better understand the aspects that help in increasing the biological activity of cruzain inhibitors, therefore helping develop effective and selective drugs to combat Chagas disease.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c03376.

SMILES structures, list of descriptors, full lists of predictions of each molecule, from training and test sets, for all models; figures with predictions × experimental and residuals; and results from the virtual screening (PDF)

Molecules and model files (ZIP)

AUTHOR INFORMATION

Corresponding Authors

Rafael Bello Gonçalves − *Department of Pharmacy, School of Pharmaceutical Sciences, University of São Paulo, São Paulo-SP 05508-900, Brazil*; Email: rafael.bello.goncalves@alumni.usp.br

Gustavo Henrique Goulart Trossini − *Department of Pharmacy, School of Pharmaceutical Sciences, University of São Paulo, São Paulo-SP 05508-900, Brazil*; orcid.org/0000-0003-3634-2531; Email: trossini@usp.br

Authors

Witor Ribeiro Ferraz − *Department of Pharmacy, School of Pharmaceutical Sciences, University of São Paulo, São Paulo-SP 05508-900, Brazil*

Raisa Ludmila Calil − *Department of Pharmacy, School of Pharmaceutical Sciences, University of São Paulo, São Paulo-SP 05508-900, Brazil*

Marcus Tullius Scotti − *Laboratory of Cheminformatics, Instituto de Pesquisa em Fármacos e Medicamentos (IPeFarM), Universidade Federal da Paraíba, Campus I, Cidade Universitária, João Pessoa 58051-900 Paraíba, Brazil*; orcid.org/0000-0003-4863-8057

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c03376

Author Contributions

All authors have read and agreed to the published version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Prof. Rino Ragno (Rome Center for Molecular Design, Department of Drug Chemistry and Technology, Sapienza Rome University, P. le A. Moro 5, 00185 Rome, Italy), who kindly set the molecule limit from www.3d-qsar.com from 50 to 70 molecules for us and answered all of our questions, making the Py-CoMFA analysis viable.

REFERENCES

(1) Oliveira, W. K.; Said, R. F. d. C.; Júnior, F. E. F. d. L.; Wada, M. Y.; Lima, M. M.; Lima, S. P.; Costa, V. M. d. Boletim Epidemiológico Doença de Chagas. *Boletim Epidemiológico*; Boletim Epidemiológico, 2020; p 45.

(2) Maguire, B. J.; Dahal, P.; Rashan, S.; Ngu, R.; Boon, A.; Forsyth, C.; Strub-Wourgaft, N.; Chatelain, E.; Barreira, F.; Sosa-Estani, S.; Guérin, P. J. The Chagas Disease Study Landscape: A Systematic Review of Clinical and Observational Antiparasitic Treatment Studies to Assess the Potential for Establishing an Individual Participant-Level Data Platform. *PLoS Negl. Trop. Dis.* **2021**, *15* (8), No. e0009697.

(3) Dias, L. C.; Dessoy, M. A.; Silva, J. J. N.; Thiemann, O. H.; Oliva, G.; Andricopulo, A. D. Quimioterapia Da Doença de Chagas: Estado Da Arte e Perspectivas No Desenvolvimento de Novos Fármacos. *Quim. Nov.* **2009**, *32* (9), 2444−2457.

(4) Ferreira, R. A. A.; Pauli, I.; Sampaio, T. S.; de Souza, M. L.; Ferreira, L. L. G.; Magalhães, L. G.; Rezende, C. d. O.; Ferreira, R. S.; Krogh, R.; Dias, L. C.; Andricopulo, A. D. Structure-Based and Molecular Modeling Studies for the Discovery of Cyclic Imides as

Reversible Cruzain Inhibitors With Potent Anti-Trypanosoma Cruzi Activity. *Front. Chem.* **2019**, *7* (November), 1−21.
(5) Rocha, D. A.; Silva, E. B.; Fortes, I. S.; Lopes, M. S.; Ferreira, R. S.; Andrade, S. F. Synthesis and Structure-Activity Relationship Studies of Cruzain and Rhodesain Inhibitors. *Eur. J. Med. Chem.* **2018**, *157*, 1426−1459.
(6) Vital, D. G.; Damasceno, F. S.; Rapado, L. N.; Silber, A. M.; Vilella, F. S.; Ferreira, R. S.; Maltarollo, V. G.; Trossini, G. H. G. Application of Bioisosterism in Design of the Semicarbazone Derivatives as Cruzain Inhibitors: A Theoretical and Experimental Study. *J. Biomol. Struct. Dyn.* **2017**, *35* (6), 1244−1259.
(7) Pauli, I. *Planejamento de Inibidores Da Enzima Cruzaína Candidatos a Fármacos Para o Tratamento Da Doença de Chagas*; Universidade de São Paulo, 2016.
(8) Sajid, M.; Robertson, S. A.; Brinen, L. S.; McKerrow, J. H. Cruzain: The Path from Target Validation to the Clinic. In *Cysteine Proteases of Pathogenic Organisms*; Dalton, M. W. R. and J. P., Ed.; Landes Bioscience and Springer Science+Business Media, 2011; Vol. *712*, pp 100−115.
(9) Trossini, G. H. G.; Guido, R. V. C.; Oliva, G.; Ferreira, E. I.; Andricopulo, A. D. Quantitative Structure-Activity Relationships for a Series of Inhibitors of Cruzain from Trypanosoma Cruzi: Molecular Modeling, CoMFA and CoMSIA Studies. *J. Mol. Graph. Model.* **2009**, *28* (1), 3−11.
(10) Pinzi, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* **2019**, *20* (18), 4331.
(11) Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10* (1), 95−115.
(12) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2019**, *18* (6), 463−477.
(13) de Souza, M. L.; Júnior, C. d. O. R.; Ferreira, R. S.; Chávez, R. M. E.; Ferreira, L. L. G.; Slafer, B. W.; Magalhaes, L. G.; Krogh, R.; Oliva, G.; Cruz, F. C.; Dias, L.; Andricopulo, A. D.Discovery of Potent, Reversible, and Competitive Cruzain Inhibitors with Trypanocidal Activity: A Structure-Based Drug Design Approach*J. Chem. Inf. Model.***2019**531028−1041
(14) Mott, B. T.; Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Ang, K. K.-H.; Leister, W.; Shen, M.; Silveira, J. T.; Doyle, P. S.; Arkin, M. R.; McKerrow, J. H.; Inglese, J.; Austin, C. P.; Thomas, C. J.; Shoichet, B. K.; Maloney, D. J. Identification and Optimization of Inhibitors of Trypanosomal Cysteine Proteases: Cruzain, Rhodesain, and TbCatB. *J. Med. Chem.* **2010**, *53* (1), 52−60.
(15) Nicklaus, M. C.; Ihlenfeldt, W.-D.; Sitzmann, M.; Filippov, I. V.; Ihlenfeldt, W.-D.; Oellien, F.; Bienfait, B.; Voigt, J. H.; Sun, G. Online SMILES Translator and Structure File Generator. https://cactus.nci.nih.gov/translate/ (accessed 04 21, 2020).
(16) Allouche, A.-R. Gabedit-A graphical user interface for computational chemistry softwares. *J. Comput. Chem.* **2010**, *32*, 174−182.
(17) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.
(18) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner BT - Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319−326.
(19) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.·.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian ∼ 09*. Revision W.04.
(20) Huang, L.; Brinen, L. S.; Ellman, J. A. Crystal Structures of Reversible Ketone-Based Inhibitors of the Cysteine Protease Cruzain. *Bioorg. Med. Chem.* **2003**, *11* (1), 21−29.
(21) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53* (13), 4891−4905.
(22) Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis. *Protein Sci.* **2018**, *27* (1), 14−25.
(23) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Sci.* **2021**, *30* (1), 70−82.
(24) Ragno, R. Www.3d-Qsar.Com: A Portal to Build 3-D QSAR Models. *Proceedings* **2019**, *22* (1), 76.
(25) Ragno, R. www.3d-qsar.com: a web portal that brings 3-D QSAR to all electronic devices—the Py-CoMFA web application as tool to build models from pre-aligned datasets. *J. Comput. Aided. Mol. Des.* **2019**, *33* (9), 855−864.
(26) Ragno, R.; Esposito, V.; Di Mario, M.; Masiello, S.; Viscovo, M.; Cramer, R. D. Teaching and Learning Computational Drug Design: Student Investigations of 3D Quantitative Structure-Activity Relationships through Web Applications. *J. Chem. Educ.* **2020**, *97* (7), 1922−1930.
(27) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, *60* (4), 1969−1982.
(28) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graph. Model.* **2002**, *20* (4), 269−276.
(29) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv J. Chim* **1980**, *4*, 359.
(30) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. *J. Mol. Biol.* **1997**, *267* (3), 727−748.
(31) Rosas-Jimenez, J. G.; Garcia-Revilla, M. A.; Madariaga-Mazon, A.; Martinez-Mayorga, K. Predictive Global Models of Cruzain Inhibitors with Large Chemical Coverage. *ACS Omega* **2021**, *6* (10), 6722−6735. Mar 16

https://doi.org/10.1021/acsomega.3c03376
ACS Omega 2023, 8, 38961−38982