A validated and interpretable predictive model of cruzain inhibitors

Jose G. Rosas-Jimenez[1,2], Marco A. Garcia-Revilla[1], Abraham Madariaga-Mazon[2], Karina Martinez-Mayorga[2*],

1 Departamento de Quimica, Division de Ciencias Naturales y Exactas, Universidad de Guanajuato, Guanajuato, Mexico

2 Instituto de Quimica, Universidad Nacional Autonoma de Mexico, Ciudad de Mexico, Mexico

* kmtzm@unam.mx

Abstract

Chagas disease affects 8–11 million people worldwide, most of them living in Latin America. Moreover, migratory phenomenon have spread the infection beyond endemic areas. Efforts for the development of new pharmacological therapies are paramount, as the pharmacological profile of the two marketed drugs currently available, nifurtimox and benznidazole, needs to be improved. Cruzain, a parasitic cysteine protease, is one of the most attractive biological targets due to its roles in parasite survival and immune evasion. In this work, we generated Quantitative Structure-Activity Relationship linear models for the prediction of pIC_{50} values of cruzain inhibitors. The statistical parameters for internal and external validation indicate high predictability with a cross-validated correlation coefficient of $q^2_{cv} = 0.77$ and an external correlation coefficient of $r^2_{ex} = 0.71$. The applicability domain is quantitatively defined, according to QSAR good practices, using the leverage method. A qualitative interpretation of the model is provided based on protein-ligand interactions obtained from docking studies and structural information codified in the molecular descriptors relevant to the QSAR

model. The model described in this work will be valuable for the discovery of novel cruzain inhibitors.

Author summary

Chagas disease is a major health problem in Latin America. The disease involves a long-lasting silent phase that usually culminates in serious or fatal heart damage. Despite its prevalence, there are only two antichagas approved drugs available. Despite these drugs have been in the market for more than 50 years, significant undesirable side effects and modest effectiveness in the chronic phase are prevalent. The need of new drugs to treat this disease is evident. Cruzain is a vital protein for the survival of *Trypanosoma cruzi*, the parasite causative of Chagas disease. Inhibition of this species-specific protein has been associated with improvements in pharmacological effects in animal models. Thus, blocking the activity of cruzain is an attractive approach for the development of antichagas agents. In this work, we present a validated mathematical model capable of predicting the cruzain inhibition value of a molecule from its chemical structure. This model can contribute to the identification of potential pharmacological alternatives against Chagas disease.

Introduction

Chagas Disease affects 8 - 11 million people in 21 Latin American countries, there is an estimation of 70 - 150 million people at risk of infection [1, 2]. Migration phenomenon have contributed to the spread of the parasite into non-endemic areas such as the United States, Europe, New Zealand, and Australia [1]. Chagas disease is a vector-borne parasitic infection caused by *Trypanosoma cruzi* and it is transmitted by the three main genera of triatomine bug, *Triatoma*, *Rhodnius*, and *Panstrongylus*. World Health Organization has recognized this infection as a Neglected Tropical Disease (NTD) because of its persistence in developing countries, being a major economic and social problem in these regions, and one of the main causes of premature death for heart failure [2–4]. It was previously reported that this disease causes an estimated loss of 752 000 working days in southern American countries [4], which implies an economic burden

of about US$1.2 billion in productivity. Globally, this parasitic infection has an estimated annual cost of $627.46 million, and 10% of this affects non-endemic countries [4]. Currently, there are only two approved drugs for the treatment of Chagas Disease: Nifurtimox (NFX) and Benznidazole (BZ). Both NFX and BZ have similar efficacy during the acute phase of infection, with 88 – 100 % of negative parasite detection after treatment with NFX and up to 80 % for BZ [5]. However, in the chronic phase, the rate of negative tests for the disease after treatment falls to 7 - 8 % [5], and there are significant side effects, including anorexia, weight loss, paresthesia, nausea, and vomiting, among others [3, 5]. Recent therapeutic research is focused on specific biological targets, which include cysteine proteases, enzymes in trypanothione metabolism, enzymes in ergosterol biosynthesis and the kinetoplastid proteasome [5].

Cruzain is a cathepsin L-like cysteine protease present in all stages of the parasite life cycle. It plays significant roles in the trypanosomal growth, survival and evasion from the host immune response. Plasma membrane-anchored cruzain degrades the Fc fraction of antibodies, overcoming the classic path of complement activation [3, 6]. In the amastigotic intracellular stage, this cysteine protease degrades transcription factors, such as NFkB and thus prevents the activation of macrophages [3]. Cruzain generates the bloodstream pro-inflammatory peptide Lys-bradykinin, which activates host immune cells, promoting the parasite uptake and spread by phagocytosis [6]. The use of cruzain inhibitors in animal models has shown to be effective in clearing the parasite burden, even in the chronic phase. The vinyl-sulphonic compound known as K777 was one the first proof-of-concept studies about anti-tripanosomal activity of cruzain inhibitors in animal models [7, 8]. Parasite death induced by cruzain inhibitors is attributed to the accumulation of a peptide precursor in the Golgi complex. Therefore, these *in vitro* and *in vivo* evidence have validated cruzain as a potential biological target for Chagas Disease [3, 6]. A variety of chemotypes for cruzain inhibition have been explored through Structure-Activity Relationships (SAR) analysis, high-throughput screening and docking methods. The most potent molecules belong to the vinyl-sulfone derivatives, oxadiazoles, nitrile-containing peptidomimetics, and thiosemicarbazones, with a broad range of biological activities among chemical families [2]. These molecules should be further optimized by increasing their selectivity towards parasite vs human cathepsins, and they should be neutral at physiological pH,

to avoid concentration in lysosomes and off-target effects [2].

Quantitative Structure-Activity Relationships (QSAR) is a ligand-based approach that mathematically correlates structural properties of molecules with their biological activity. QSAR modeling is widely used in drug discovery, especially in the prediction of enzyme inhibition and ADME-Tox properties [9]. In virtual screening, validated QSAR models are used for prioritizing molecules for experimental evaluation. Carefully validated QSAR models have rendered novel chemotypes and scaffolds with a desirable biological activity [10]. The quality of a QSAR model can be evaluated using the OECD principles [11]. These principles are a series of guidelines originally developed for the use of QSAR modelling for regulatory purposes, but they became a valuable tool in the standard QSAR practice [11, 12]. In this work, we explored public databases of structurally diverse cruzain inhibitors for the generation of QSAR predictive models of this biological endpoint. The structural properties, encoded by molecular descriptors, are rationalized in terms of protein-inhibitor interactions, using molecular docking, thus providing a possible mechanistic interpretation of the model. This work will be useful in the search of cruzain inhibitors.

Materials and methods

Data compilation, curation, and pre-processing

Cruzain inhibitors were collected from the ChEMBL (release 24) database, searching by molecular target using the keyword *cruzain*. Molecules annotated with IC_{50} values were selected and duplicated or missing values compounds were eliminated. Finally, a selection based on the same experimental protocol for IC_{50} determination was performed. The selected experimental procedure is a competitive fluorescence assay in the presence of detergent, as reported by Babaoglu *et al* [13]. The detergent is used to avoid aggregation, which is the main cause of false positives in exploratory and high throughput screens [2, 14]. Structural and biological information of the compounds was verified in the corresponding original publications, and when required, the discrepancies were fixed. IC_{50} values were transformed to pIC_{50}. The final dataset consisted of 110 structurally diverse cruzain inhibitors. The 2D and 3D coordinates of these molecules

were calculated from their SMILES representation, using the *wash* tool in the software Molecular Operating Environment 2019.01 (MOE) [15]. Lastly, the structures were energy minimized using the MMFF94x force field. The curated database is available in S1 Table.

To summarize the chemical diversity in the set, the MACCSKeys fingerprints as implemented in RDKit [16] package were calculated for every molecule. A clustering calculation was performed using the affinity propagation algorithm, with the Tanimoto similarity matrix as affinity measure. The chemical structures for the representative molecules in every cluster are presented below.

QSAR modeling

Descriptor calculation and feature selection

All molecular descriptors available in MOE were calculated, including topostructural and topochemical indices, subdivided Van der Waals surface areas, VolSurf potentials, and physicochemical properties such as dipolar and hydrophobic moments. The dataset was randomly split into a training set (88 molecules, 80%) and a test set (22 molecules, 20%). Descriptors were scaled to [0,1] range using Eq (1) and those in the test set were scaled according to the training set. Constant descriptors (zero variance) were filtered out.

$$X_i' = \frac{X_i - X_{i,min}}{X_{i,max} - X_{i,min}} \tag{1}$$

Feature selection and model calculation were performed in Weka 3.8 [17,18]. Selection of relevant features was carried out using the Correlation-Based Feature Selection (CFS) with a Greedy Stepwise algorithm [19]. Briefly, CFS calculates a merit score, M_s, on a subset of variables through Eq (2), where r_{fc} is the average pair-wise correlation coefficient between features and the dependent variable, r_{ff} is the average pair-wise correlation coefficient between features themselves, and k is the number of features. Higher merit values involve a higher correlation with the dependent variable and a less correlation between features, penalizing high-dimensional sets also. In the Greedy Stepwise algorithm, the variables are sequentially added until the merit reaches a maximum.

$$M_s = \frac{k r_{fc}}{\sqrt{k + k\,(k-1)\,r_{ff}}} \tag{2}$$

The subset of features with the highest score were used in the generation of the Multiple Linear Regression model, as implemented in Weka 3.8.

Model validation

The goodness of fit for the model was estimated by calculating the following statistical parameters: coefficient of determination (R^2), adjusted coefficient of determination ($R^2 - adj$), F statistic (variance ratio) and its associated p-value. Internal validation was carried out through the k-fold leave-some-out cross-validation with $k = 10$. The cross-validated correlation coefficient (q^2_{cv}) is reported to evaluate the robustness of the model. Model predictability was assessed by applying the generated equation to calculate the biological data of the test set. Using these results, the Golbraikh and Tropsha (G&T) external validation parameters were calculated in the Enalos nodes for Knime [20–22]. Golbraikh and Tropsha parameters use regression through the origin to estimate the deviation of the model with respect to the ideal QSAR regression. Basically, these parameters compare the differences of the coefficients of determination and slopes of the fitted model and the regression forced to the origin, R_0 and k. The model is considered predictive if all parameters are within defined thresholds [20].

Applicability Domain

The predictivity of a QSAR model is framed by the nature of the molecules in the training set. The applicability domain is the quantitative delimitation of the chemical space where predictions are reliable. In this work, the applicability domain was defined using the leverage method [23]. Leverage values, h_i, are computed using Eq (3), where $\mathbf{X}$ is the descriptor matrix of the training set and $\mathbf{x_i}$ is the descriptor vector for a query molecule.

$$h_i = \mathbf{x}_i^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_i \tag{3}$$

Basically, leverage values are proportional to the distance of the molecule from the centroid of the training set. Thus, compounds above a threshold are far from the

explored chemical space and therefore, their predicted biological activity will be unreliable. Typically, the threshold, h_{max}, is computed with Eq (4), where p is the number of features and n is the number of molecules in the training set.

$$h_{max} = 3\frac{p}{n} \tag{4}$$

Leverage and limit values were computed with the Applicability Domain node calculator of Enalos for Knime [24, 25]. Results are presented in a Williams plot (leverage vs standarized residuals), where outliers in the activity domain or structurally influential, can be visually detected. The Williams plot is a representation of the chemical space spanned by the model.

Docking calculation

The coordinates of cruzain were downloaded from the Protein Data Bank with PDB-ID code 3KKU [26]. This structure is reported with a resolution of 1.28 Å, and it is co-crystalized with a benzimidazole derived, a non-covalent ligand. The protein was prepared in MOE as follows: hydrogen atoms were added according to protonation states at pH 7.0 and Gasteiger-Marsili charges were computed [27]. The protein structure file was converted to the PDBQT format. Gasteiger charges were also computed for the ligands and they were converted to PDBQT format. The docking calculation was performed in AutoDock Vina [28]. The search space was extended in the binding site of the cruzain with a box of size 24 Å $*$ 30 Å $*$ 20 Å. The docking calculation was performed in 10 repetitions, and the conformations with the best score per molecule were selected to generate a database of bound conformations. These data were used to generate protein-ligand interaction fingerprints in MOE.

The similarity maps tool included in the RDKit module for python was used to generate partial charges and SLogP diagrams. These diagrams show the atomic contributions to logP, calculated with the Wildman and Crippen algorithm [29], and the Gasteiger partial charges [27]. The 2D depictions of molecules in the similarity maps were generated by projecting the calculated 3D conformation of the molecules, so that these depictions resemble their docked pose.

Results and Discussion

In this work, we present the preparation and analysis of a data set of 110 cruzain inhibitors annotated with pIC_{50} values. The distributions of the biological activity values of the training and test sets are shown in Fig 1. The pIC_{50} values ranges from 3.48 to 10.0 units, from nanomolar to sub-millimolar scale. Typically, the reported experimental error for this biological assay is around 2.0 μM, which implies that the range of pIC_{50} values of this dataset is more than five times higher than the experimental error, in agreement with general recommendations and best practices for QSAR modeling [10]. Noteworthy, the biological activity values of the test set lie within that of the training set as shown in the histogram of the Fig 1 and with no gaps within bins.

Fig 1. Distribution of pIC_{50} values of 110 cruzain inhibitors. Molecules in the training set (88) are shown in gray, and 22 molecules in the test set are shown in black. The inhibitory potency of the test set fall within interval of pIC_{50} values of the training set.

To summarize the chemical diversity contained in the dataset, we performed a clustering analysis with the affinity propagation algorithm, using Tanimoto similarity from MACCSKeys fingerprints as affinity measure, in the RDKit and Sci-Kit learn modules in Python. Fig 2 depicts the calculated clusters, along with the representative structure from every cluster. The chemical families include thiosemicarbazones, acylhydrazines, oxadiazoles and nitrile-containing peptidomimetics. Thiosemicarbazones are the most numerous compounds in the set and peptidomimetics are the most potent known inhibitors, as has been reported previously [2].

Fig 2. Representation of the molecular diversity in the dataset. Molecules are grouped according to the clustering results by affinity propagation. Structures for representative molecules, highlighted with a cross-shaped mark, are shown for every cluster.

The MLR algorithm generates an explicit equation, consisting of a linear combination of molecular descriptors. After selecting the feature subset which renders the maximum merit score, as described in the methods section, Eq 5 was obtained for the estimation of pIC_{50} values. A brief definition of the descriptors involved in the model is presented in Table 1. In general, these descriptors account for electrostatic

(E_{ele} and PEOE descriptors), hydrophobic (SLOGP and vsurf_ID8 descriptors) and hydrophilic (vsurf_W related descriptors) properties. These features are crucial for the establishment of potential intermolecular interactions required for the binding of ligands into the active site. Therefore, the linear equation may be related to the presence of such features in the binding process.

$$-log\,(IC_{50}) = pIC_{50} = -1.30a\_nF + 3.62E_{ele} + 2.34GCUT\_SLOGP2 + 2.46PEOE\_VSA(-1) + 2.18PEOE\_VSA(-3) - 1.24SLOGP\_VSA4 + 0.69SLOGP\_VSA9 - 1.07vsurf\_DW12 - 1.56vsurf\_EWmin1 + 1.69vsurf\_ID8 + 1.21vsurf\_Wp5 + 1.06 \quad (5)$$

Table 1. Definition of molecular descriptors selected in the linear equation of the model.

Descriptor	Definition
a_nF	Number of fluorine atoms
E_{ele}	Electrostatic component of potential energy
GCUT_SLOGP_2	The GCUT descriptors using atomic contribution to logP using the Wildman and Crippen SlogP method.
PEOE_VSA_-1	Sum of v_i where q_i is in the range $[-0.10, -0.05)$.
PEOE_VSA-3	Sum of v_i where q_i is in the range $[-0.20, -0.15)$.
SLOGP_VSA4	Sum of v_i such that L_i is in $[0.1, 0.15)$.
SLOGP_VSA9	Sum of v_i such that $L_i > 0.40$
vsurf_DW12	Contact distances of vsurf_EWmin
vsurf_EWmin1	Lowest hydrophilic energy
vsurf_ID8	Hydrophobic integy moment (-1.6 kcal/mol)
vsurf_Wp5	Polar volume (-3.0 kcal/mol)

v_i is the atomic Van der Waals surface area of atom i, q_i is the Gasteiger-Marsili partial charge over the atom i, and L_i is the Wildman-Crippen atomic contribution to LogP.

Statistical parameters describing the goodness of fit for the model are presented in Table 2. Coefficients of determination near to 1 indicates that a high ratio of variance present in the original data is explained by the model. In this case, 83% of the variance already present in the pIC_{50} of the training set is explained by Eq 5. The ratio of the mean squared error of the one-parameter model and the generated model is measured by the F statistic. If this ratio is high enough, the prediction made by Eq 5 has an error less than the native variability in the data. The F value for this model is presented in Table 2 along with its associated $p - value$. Making the assumptions of the linear model, the probability of finding an F ratio of 34.08 or higher, for a 10 parameter equation, is less than 0.001, if the model error is equal to the variability in data. Given

this low probability value, this hypothesis can be rejected and accept that predictions made by the model equation are more accurate than just the mean value around the standard deviation for the original data.

Table 2. Statistical parameters describing the goodness of fit for the model.

Parameter	Value
R^2	0.83
R^2-adjusted	0.81
F ratio	34.08
p-value	< 0.001

Since conclusions derived from statistical parameters rely on the parametric assumptions, their fulfillment were tested by means of an analysis of residuals, shown in Fig (3). In the linear model, the dependent variable, Y_i, has a normal distribution around the predicted value $\hat{Y}_i$, thus the prediction error, $Y_i - \hat{Y}_i$, must follow a normal distribution with a mean of 0. The lower panel of Fig (3) shows the quantile plot for the calculated errors and the theoretical normal distribution. Most of the values in the quantile plot follow a straight line, suggesting a very near behavior to a normal distribution, achieving the normality requirement.

Fig 3. Histograms with the distibution of residuals, as predicted with Eq 5. The quantile plots, comparing to a normal distribution are also presented, for both the training and the test set. The regression line shows a near behavior to a normal distribution.

Observed and predicted activity values for both training and test sets are shown in Fig (4). The pIC_{50} values for the training set were calculated in a 10-fold cross validation step, thus the coefficient of determination in Fig (4) corresponds to $Q^2 - LSO$. The test set, not used for the model construction, has a clear behavior near to the linear fit. The R^2 for this external set is 0.71, above the typical threshold of 0.6. However, although a high value of both Q^2 and R^2 is required, it is not sufficient for the predictability estimation since these parameters just measure the linear correspondence between predicted and experimental values but not their 1:1 identity relationship [20]. Since there is not consensus in the establishment of an universal predictability criteria for QSAR modeling, one of the proposed practices is to calculate a set of parameters that could characterize the deviation from an ideal prediction, as suggested by Chirico *et al* [30] and Gramatica *et al* [31].

Fig 4. Regression plot for the results of predicted pIC_{50} values. The training set values shown were obtained in a 10-fold cross validation step. The coefficients of determination for both sets ($Q^2 - LSO$ and $R^2 - ext$) are also presented. Continuous and dashed gray lines are the linear fits for training and test set, respectively.

The G&T criteria measure the agreement between experimental and predicted values [20, 30, 31]. These validation parameters were developed following the idea that the regression line for a predictive model should be the identity relation, $Y_i = \hat{Y}_i$. Thus, the values of the G&T criteria measure the deviation of the least-squares line for the model from the identity straight line. Table 3 shows the results of these criteria, for the external evaluation used in this work, along with their acceptance thresholds as suggested by the authors. All the values are within the acceptable range, indicating a good agreement between the experimental information and the predictions of the model using the external test set.

Table 3. Golbraikh and Tropsha parameters and criteria for external validation calculated for the model

G&T Criterion	Value
$R^2 > 0.6$	0.71
$R^2_{cvext} > 0.5$	0.66
$(R^2 - R_0^2)/R^2 < 0.1$	0.05
$(R^2 - R_0^{'2})/R^2 < 0.1$	0.02
$ABS(R_0^2 - R_0^{'2}) < 0.1$	0.02
$0.85 < k < 1.15$	0.95
$0.85 < k' < 1.15$	1.03

Applicability domain was defined using the leverage method, using both the training set and the test set. Williams plot for the dataset is presented in Fig (5). Because leverage is a projection of the distance from the training set, the distribution of the molecules in the Williams plot is a representation of the chemical space covered by the model. Standardized residuals are distributed around the expected value of 0, as was shown previously, for both the training and test sets. It is interesting to note that most of the test molecules follow a distribution similar to those in the training set, and their residuals are inside the expected errors predicted for the training set. It is also remarkable that two molecules in the training set and three in the test set display leverage values higher than the calculated limit. In these regions, any prediction made by the model is considered an extrapolation and its reliability is low.

Most of the molecular descriptors shown in Table 1 are related to potential

Fig 5. Williams plot for the applicability domain definition, using the leverage method.

intermolecular interactions. To rationalize the binding recognition process of cruzain inhibitors, based on the analysis of the molecular descriptors obtained in the QSAR model, molecular docking simulations were performed. PLIF histograms in Fig (6) summarize these results from the database of bound conformations. Fig (6A) shows interactions involving atom pairs between the protein and the ligand, whereas Fig (6B) summarizes surface contact interactions. These histograms show that hydrogen bond formation and polar contacts are predominant in the S1 subsite and near the catalytic site, whereas in S2 and S1' subsites, hydrophobic contacts and π interactions are more favorable. Regarding with such interactions, molecular descriptor vsurf_ID8 is the hydrophobic integy moment (INTEraction enerGY) at -1.6 kcal/mol as defined by Cruciani *et al* [32]. Basically, the hydrophobic integy moment is the unbalance between the center of mass of the molecules and the hydrophobic regions. Thus, the descriptor may be related to the complementariety of inhibitors with the binding site, i.e. the ability to form hydrogen bonds or polar contacts with the catalytic site or the S1 subsite and hydrophobic or π interactions in the S2 or S1' subsites.

Fig 6. PLIF results for the docking calculation of the cruzain inhibitors in the dataset. A: PLIF histogram for potential contacts. The color of the bars represents the binding subsite in the cruzain. The code on the top of the bar is the kind of interaction: D, A, side chain hydrogen bond donors or acceptors; d, a, backbone hydrogen bond donors or acceptors, and R, arene or π interactions. B: PLIF histogram for surface contacts. The color of the bars represents the binding subsite in the cruzain. The code on the top of the bar is the kind of surface contact: H, hydrophobic; P, partial hydrophobic; Q, charged; X, other, and C total

Subdivided Van der Waals surface area descriptors are defined in terms of properties which can be divided into atomic contributions. In this case, partial charges and logP contributions take into consideration the total available surface area for certain types of electrostatic and hydrophobic contacts. Fig (7) shows the predicted conformations for some of the molecules in the set, along with their 2D representation depicting the partial charges and the atomic contributions to logP. PEOE_VSA_-1 and PEOE_VSA_-3 account the total surface area for atoms whose partial charges are in the ranges $[-0.10, -0.05)$ and $[-0.20, -0.15)$, respectively. Atoms with partial charges related to PEOE_VSA_-1 are often carbon atoms in aromatic rings and saturated chains. These

molecular fragments bind to hydrophobic cavities, mainly in S2 and S1' subsites and remarkably they have close contacts with TRP-184. On the other hand, partial charges accounted by PEOE_VSA_-3 are related to nitrogen-nitrogen containing groups, such as thiosemicarbazones, acylhydrazines and oxadiazoles. This partial charge is also associated with nitrile nitrogen, which is a chemical group present in the peptidomimetics, the most active compounds in the set. All these groups are frequently used as mimetics of the peptide bond since they can exert polar interactions required for the backbone recognition near the catalytic and S1 subsites.

Fig 7. Binding conformations predicted by docking and visualization of descriptors related to partial charges and logP contributions. 2D depictions were generated as projections of their 3D conformations. The lines inside color bars are the ranges which contribute to the binned Van der Waals surface area descriptors in the QSAR model. In the 3D representation, cruzain subsites are shown in colors: yellow for the catalytic triad, red for S1 subsite, raspberry for the S2 subsite, deepsalmon for S3 subsite, tv_blue for S1' subsite and lightblue for S2' subsite (colors as defined by Pymol).

Regarding with surface descriptors, SLOGP_VSA4 and SLOGP_VSA9 measure the total surface area for logP atomic contributions in the ranges $[0.1, 0.15)$ and > 0.4, respectively. Most of the atomic fragments related to SLOGP_VSA4 are oxygen atoms in carbonyl groups directly attached to aromatic rings. However, the coefficient for this descriptor in the model equation is negative, indicating that this feature is unfavorable for biological activity. The last of the subdivided Van der Waals surface area descriptor takes into consideration mostly halogen atoms bound to aromatic or aliphatic groups. The most potent compounds in the data set are also rich in halogen-containing groups. Halogenated substituents are frequently used, among with other effects, to fulfill steric contacts into protein cavities, so they can exert a shape-complementary effect with the cruzain binding site, particularly in the well-defined S2 cavity and in the clefts formed by the S1 and S1' subsites.

Volsurf descriptors are calculated from grids extended around the molecule, and then computing the interaction energy of this molecule with a probe on each of the grid points. DW12, EWmin1 and Wp5 are calculated using a water molecule as a probe, and thus are representative of polar interactions. For an energy isovalue of -3.0 kcal/mol, the field is representative of favorable polar and hydrogen bond donor-acceptor regions [32]. The total polar volume at this energy (Wp5) is positively correlated with biological

activity, as can be deduced from its coefficient in the model equation. Futhermore, EWmin1 indicates that a lowest hydrophilic interaction energy is more favorable for cruzain inhibition. Fig (8) shows isosurfaces for the interaction fields at a level of -3.0 kcal/mol, with the same molecules as in Fig (7). It is clear from these representations that polar volumes extend around hydrogen bond donors and acceptors, mainly on those groups that mimic the peptide bond. Thus, these grid-based descriptors account for the ability of inhibitors to form hydrogen bonds in the binding site for the peptide bond recognition.

Fig 8. Polar surfaces at an isovalue of -3.0 kcal/mol. These interaction grids are calculated using a water molecule in every point and are indicative of polar interactions.

The interpretation provided above is based on the physical meaning of descriptors in terms of protein-ligand interactions. The model equation summarizes the presence of chemical fragments whose atoms meet the electrostatic and hydrophobic requirements for the binding into cruzain subsites but also their spatial distribution, as described by their integy moments and polar molecular fields. These requirements resemble a pharmacophore model that molecules within the applicability domain must meet to bind into the protein and exert its inhibitory effect.

In summary, we have presented a QSAR model with a well-defined endpoint, as described in methodology section for the criteria of data selection. The algorithm is unambiguously presented, which consists in the application of Eq 5 to calculate the predicted pIC_{50} for cruzain inhibition, given the required descriptors. The applicability domain is defined using the leverage method, and a limit value is also given for the reliability of predictions. The statistics for the goodness-of-fit, robustness and predictability were calculated and all of them fall within the generally accepted thresholds. Finally, a possible mechanistic interpretation of the model is proposed, in terms of intermolecular interactions. Thus, in this study, the five OECD principles for good practices in QSAR modeling are fulfilled. These principles are quality standards for QSAR developement, mainly in regulatory purposes. Under these criteria, our QSAR model is predictive and could be used in the search of new inhibitors or in the rational design of new compounds with this biological activity.

Conclusion

A Quantitative Structure-Activity Relationship model was developed for the calculation of pIC_{50} values of cruzain inhibitors using multiple linear regression. The statistical parameters describing its performance agree with the general recommendations for QSAR modeling. In particular, the external validation demonstrates high predictability, since the calculated statistical parameters are above the recommended thresholds, considering its applicability domain. The molecular descriptors selected in the model equation are related to the potential formation of intermolecular interactions as shown in the binding modes calculated by docking. The linear equation integrates partial charge, hydrophobic potentials, and energy with spatial distribution and volume availability for polar interactions, indicating that there is a pharmacophoric-like recognition in the core of this QSAR model. The use and interpretation of this model could guide in the search, development and rational design of cruzain inhibitors as possible pharmacological treatment of Chagas disease.

Supporting information

S1 Table. Data set of cruzain inhibitors. Table with the cruzain inhibitors, incluiding SMILES representation, activity values (pIC_{50}), and calculated descriptors.

Acknowledgments

K.M.-M. thanks DGAPA-UNAM (PAPIIT IN210518) and Instituto de Química, UNAM for finantial support. J.G.R.-J. thanks Biosen Institute for scholarship. Authors thank AutoDock Vina, Weka, Knime, RDKit, and Sci-Kit Learn developers for making machine learning and chemoinformatics tools freely available for academic purposes.

References

1. Flores-Ferrer A, Marcou O, Waleckx E, Dumonteil E, Gourbière S. Evolutionary ecology of Chagas disease; what do we know and what do we need? Evolutionary Applications. 2018;11(4):470–487. doi:10.1111/eva.12582.

2. Martinez-Mayorga K, Byler KG, Ramirez-Hernandez AI, Terrazas-Alvares DE. Cruzain inhibitors: efforts made, current leads and a structural outlook of new hits. Drug Discovery Today. 2015;20(7):890–898. doi:10.1016/J.DRUDIS.2015.02.004.

3. Ferreira LG, Andricopulo AD. Targeting cysteine proteases in trypanosomatid disease drug discovery. Pharmacology & Therapeutics. 2017;180:49–61. doi:10.1016/J.PHARMTHERA.2017.06.004.

4. Pérez-Molina JA, Molina I. Chagas disease. The Lancet. 2018;391(10115):82–94. doi:10.1016/S0140-6736(17)31612-4.

5. Carneiro CM, Sánchez-Montalvá A, Corrêa-Oliveira R, Sales Junior PA, Fonseca Murta SM, Salvador F, et al. Experimental and Clinical Treatment of Chagas Disease: A Review. The American Journal of Tropical Medicine and Hygiene. 2017;97(5):1289–1303. doi:10.4269/ajtmh.16-0761.

6. Sajid M, Robertson SA, Brinen LS, McKerrow JH. Cruzain. Springer, Boston, MA; 2011. p. 100–115. Available from: http://link.springer.com/10.1007/978-1-4419-8414-2{_}7.

7. Engel JC, Doyle PS, Hsieh I, McKerrow JH. Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection. Journal of Experimental Medicine. 1998;188(4):725–734. doi:10.1084/jem.188.4.725.

8. Palmer JT, Rasnick D, Klaus JL, Brömme D. Vinyl Sulfones as Mechanism-Based Cysteine Protease Inhibitors. Journal of Medicinal Chemistry. 1995;38(17):3193–3196. doi:10.1021/jm00017a002.

9. Gini G. QSAR: What Else? Humana Press, New York, NY; 2018. p. 79–105. Available from: http://link.springer.com/10.1007/978-1-4939-7899-1{_}3.

10. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics. 2010;29(6-7):476–488. doi:10.1002/minf.201000061.

11. Gómez-Jiménez G, Gonzalez-Ponce K, Castillo-Pazos DJ, Madariaga-Mazon A, Barroso-Flores J, Cortes-Guzman F, et al. The OECD Principles for (Q)SAR Models in the Context of Knowledge Discovery in Databases (KDD). In: Advances in Protein Chemistry and Structural Biology. vol. 113. Academic Press Inc.; 2018. p. 85–117.

12. Gramatica P. Principles of QSAR Modeling. International Journal of Quantitative Structure-Property Relationships. 2020;5(3):1–37. doi:10.4018/IJQSPR.20200701.oa1.

13. Babaoglu K, Simconov A, Irwin JJ, Nelson ME, Feng B, Thomas CJ, et al. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against β-lactamase. Journal of Medicinal Chemistry. 2008;51(8):2502–2511. doi:10.1021/jm701500e.

14. Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, et al. An Aggregation Advisor for Ligand Discovery. Journal of Medicinal Chemistry. 2015;58(17):7076–7087. doi:10.1021/acs.jmedchem.5b01105.

15. ULC CCG. Molecular Operating Environment (MOE); 2019.

16. RDKit: Open-source cheminformatics; 2020. Available from: `https://www.rdkit.org/`.

17. Frank E, Hall MA, Witten IH, Kaufmann M. WEKA Workbench Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"; 2016. Available from: `https://www.cs.waikato.ac.nz/ml/weka/Witten{_}et{_}al{_}2016{_}appendix.pdf`.

18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update; 2009. Available from: `https://www.kdd.org/exploration{_}files/p2V11n1.pdf`.

19. Hall MA. Correlation-based Feature Selection for Machine Learning. The University of Waikato; 1999. Available from: `https://www.cs.waikato.ac.nz/{~}mhall/thesis.pdf`.

20. Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling. 2002;20(4):269–276. doi:10.1016/S1093-3263(01)00123-1.

21. Melagraki G, Afantitis A. Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. Chemometrics and Intelligent Laboratory Systems. 2013;123:9–14. doi:10.1016/J.CHEMOLAB.2013.02.003.

22. Vrontaki E, Melagraki G, Mavromoustakos T, Afantitis A. Searching for anthranilic acid-based thumb pocket 2 HCV NS5B polymerase inhibitors through a combination of molecular docking, 3D-QSAR and virtual screening. Journal of Enzyme Inhibition and Medicinal Chemistry. 2016;31(1):38–52. doi:10.3109/14756366.2014.1003925.

23. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R, et al. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules. 2012;17(5):4791–4810. doi:10.3390/molecules17054791.

24. Afantitis A, Melagraki G, Sarimveis H, Koutentis P, Markopoulos J, Igglessi-Markopoulou O. Development and Evaluation of a QSPR Model for the Prediction of Diamagnetic Susceptibility. QSAR & Combinatorial Science. 2008;27(4):432–436. doi:10.1002/qsar.200730083.

25. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Kollias G, Igglessi-Markopoulou O. Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors. Molecular Diversity. 2009;13(3):301–311. doi:10.1007/s11030-009-9115-2.

26. Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, Keiser MJ, et al. Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. Journal of Medicinal Chemistry. 2010;53(13):4891–4905. doi:10.1021/jm100488w.

27. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron. 1980;36(22):3219–3228. doi:10.1016/0040-4020(80)80168-2.

28. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2009;31(2):NA–NA. doi:10.1002/jcc.21334.

29. Wildman SA, Crippen GM. Prediction of Physicochemical Parameters by Atomic Contributions. Journal of Chemical Information and Computer Sciences. 1999;39(5):868–873. doi:10.1021/ci990307l.

30. Chirico N, Gramatica P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. Journal of Chemical Information and Modeling. 2011;51(9):2320–2335. doi:10.1021/ci200211n.

31. Gramatica P, Sangion A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. Journal of Chemical Information and Modeling. 2016;56(6):1127–1131. doi:10.1021/acs.jcim.6b00088.

32. Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. Journal of Molecular Structure: THEOCHEM. 2000;503(1-2):17–30. doi:10.1016/S0166-1280(99)00360-7.

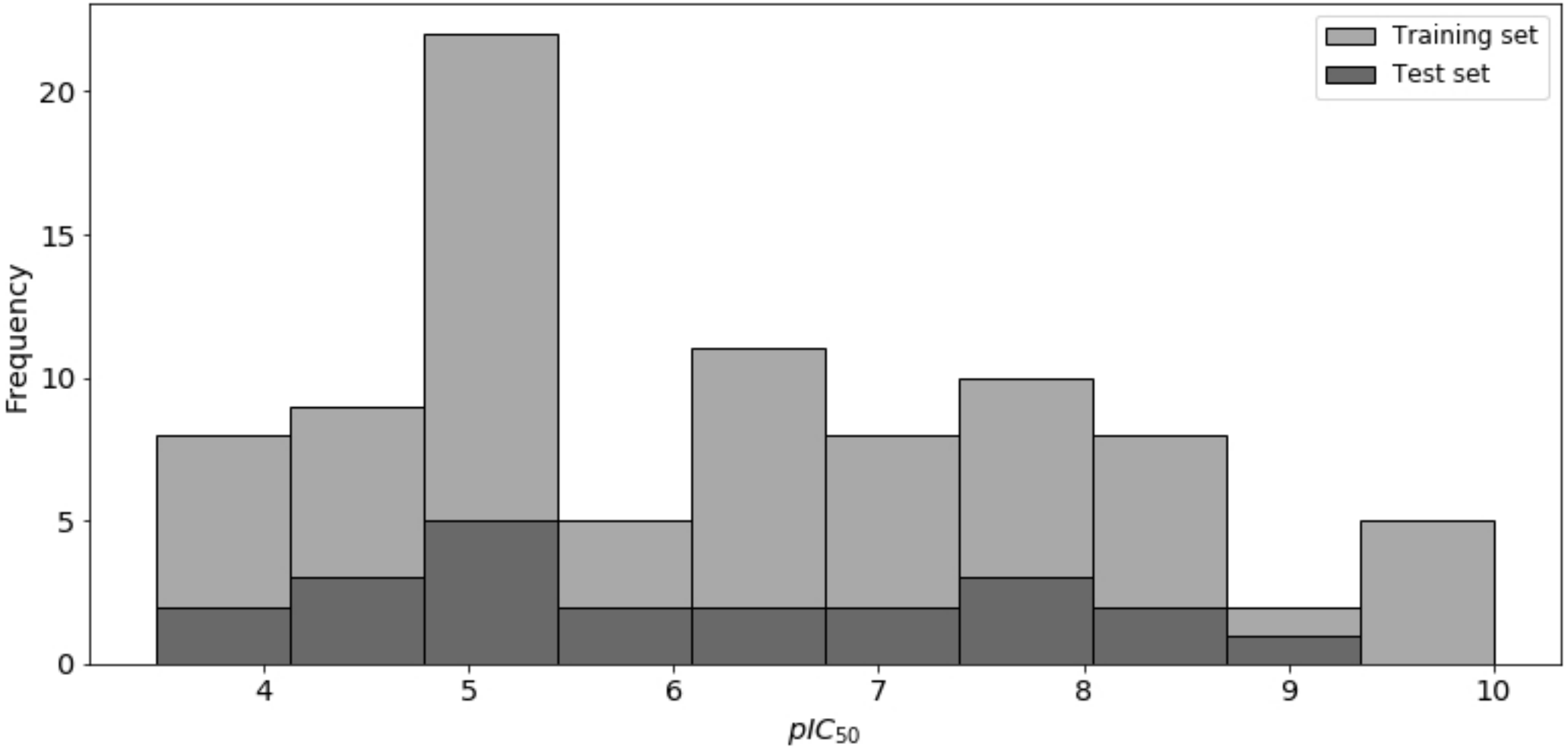


figure_1

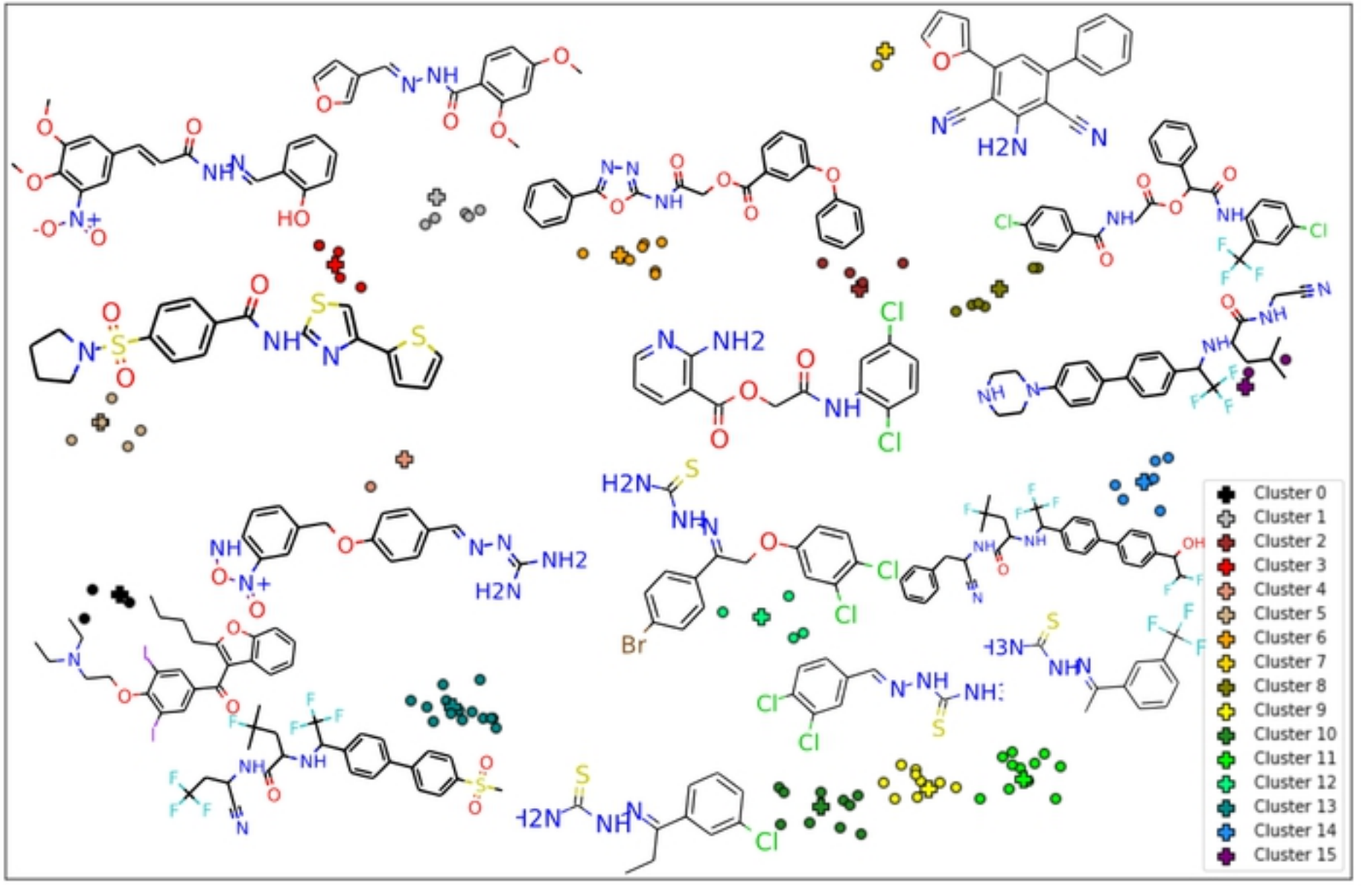

Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15

figure_2

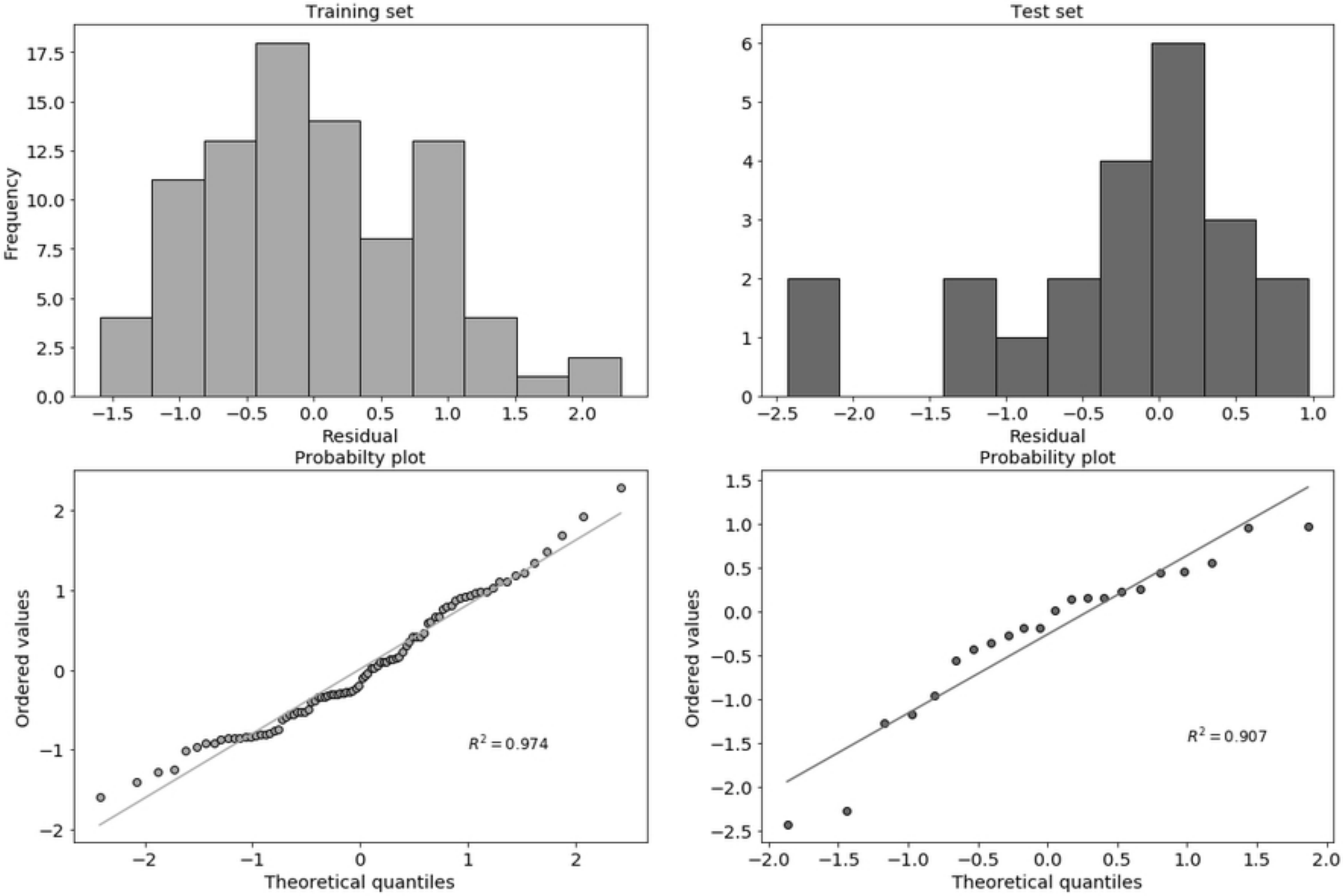

Training set
Frequency
Residual
Test set
Residual
Probabilty plot
Ordered values
Theoretical quantiles
$R^2 = 0.974$
Probability plot
Ordered values
Theoretical quantiles
$R^2 = 0.907$

figure_3

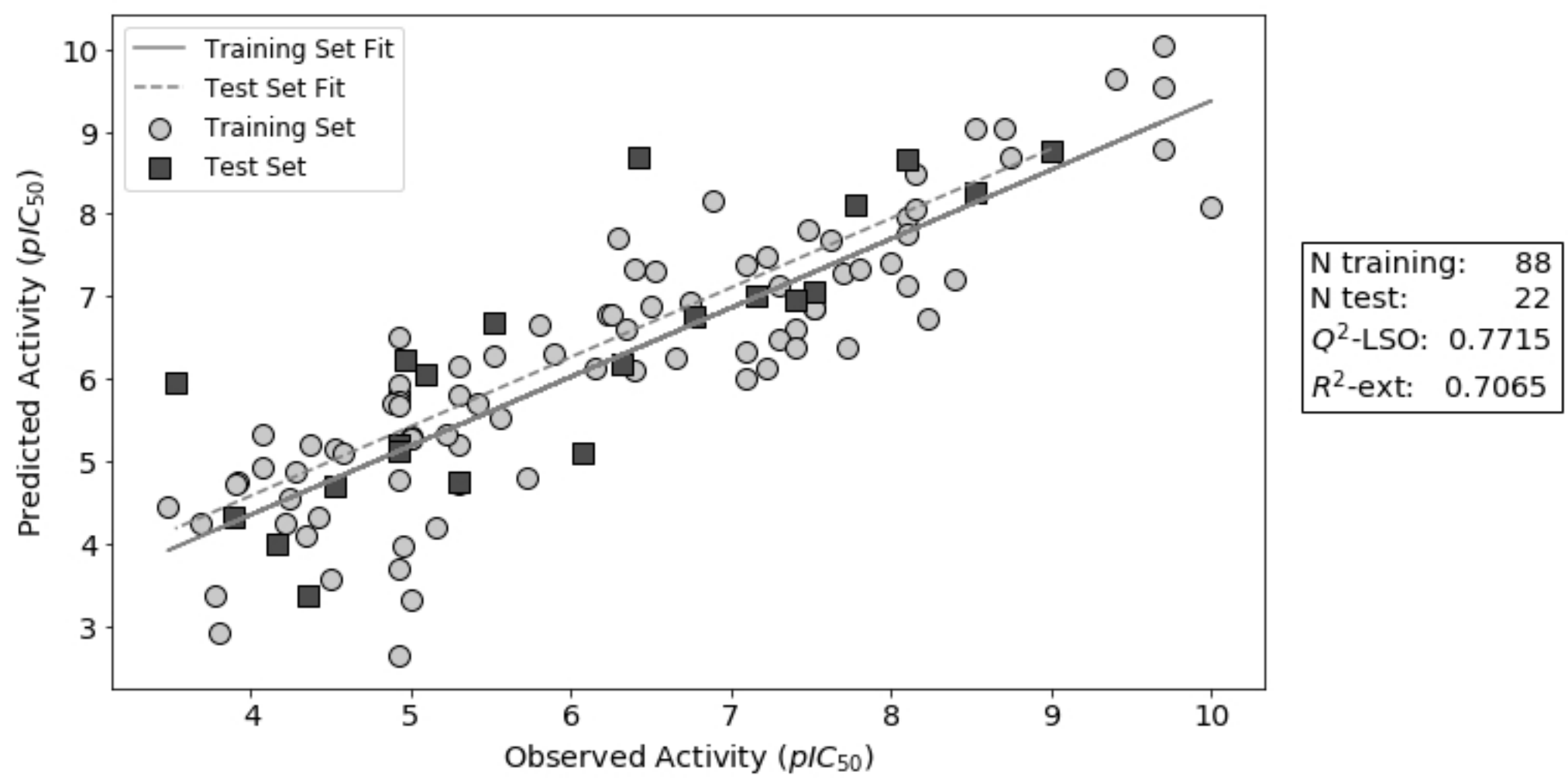

Training Set Fit
Test Set Fit
Training Set
Test Set
Predicted Activity (pIC_{50})
Observed Activity (pIC_{50})
N training: 88
N test: 22
Q^2-LSO: 0.7715
R^2-ext: 0.7065

figure_4

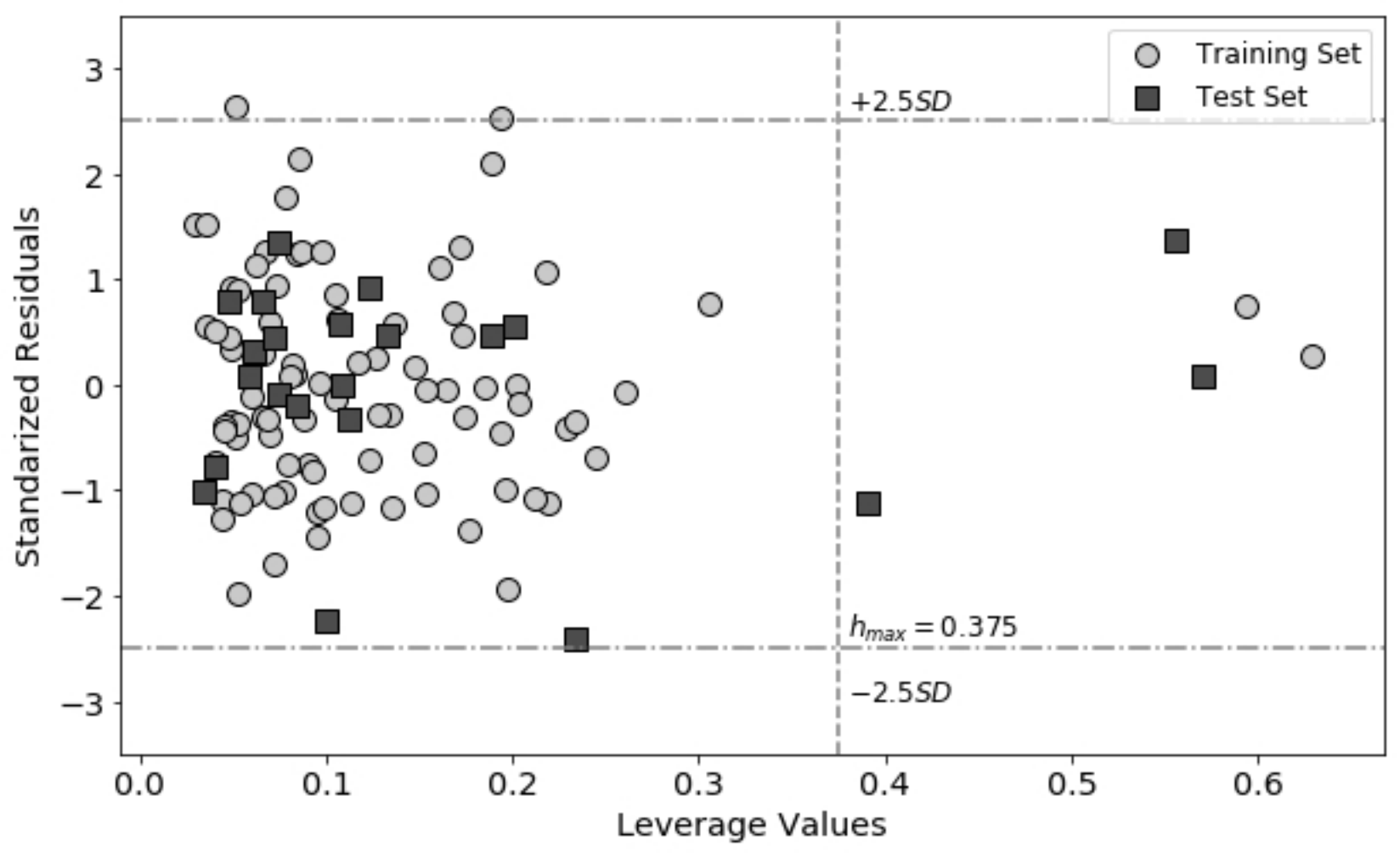
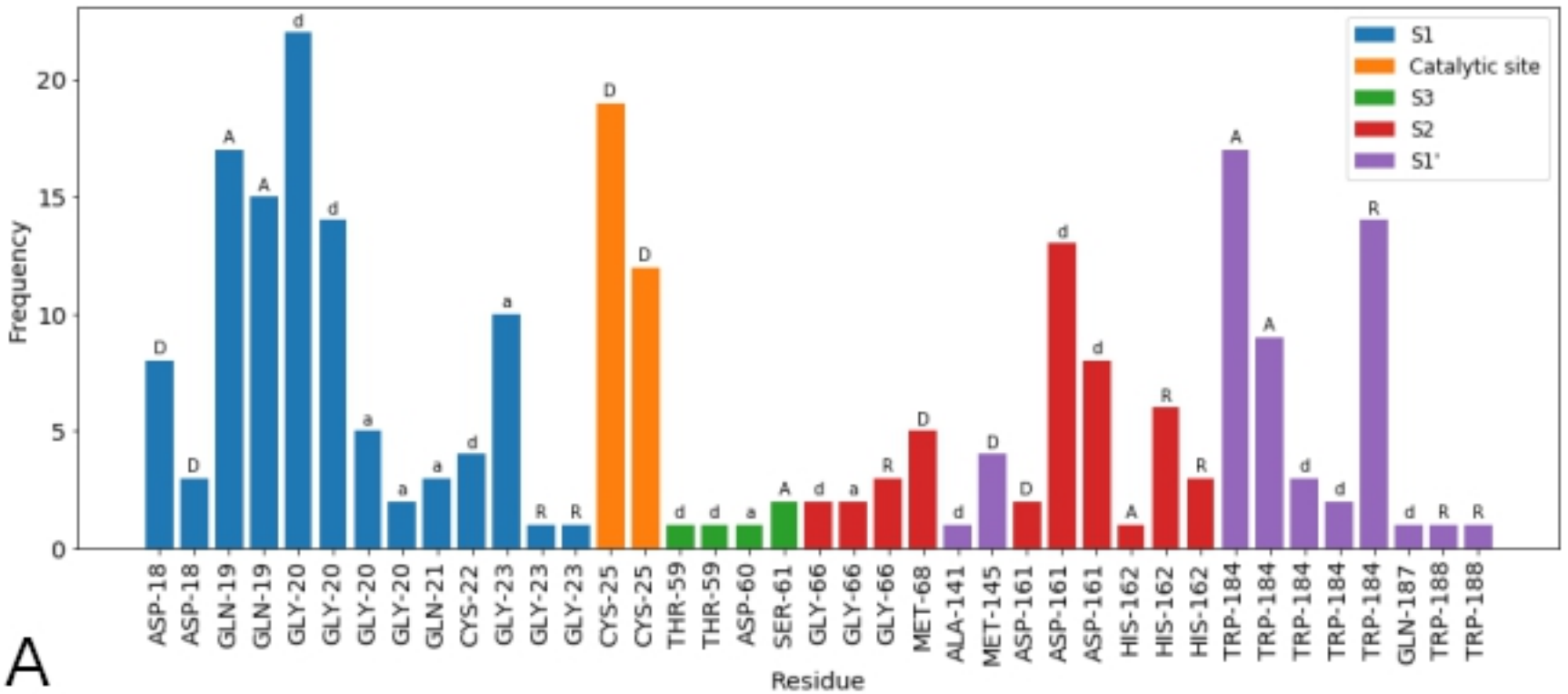
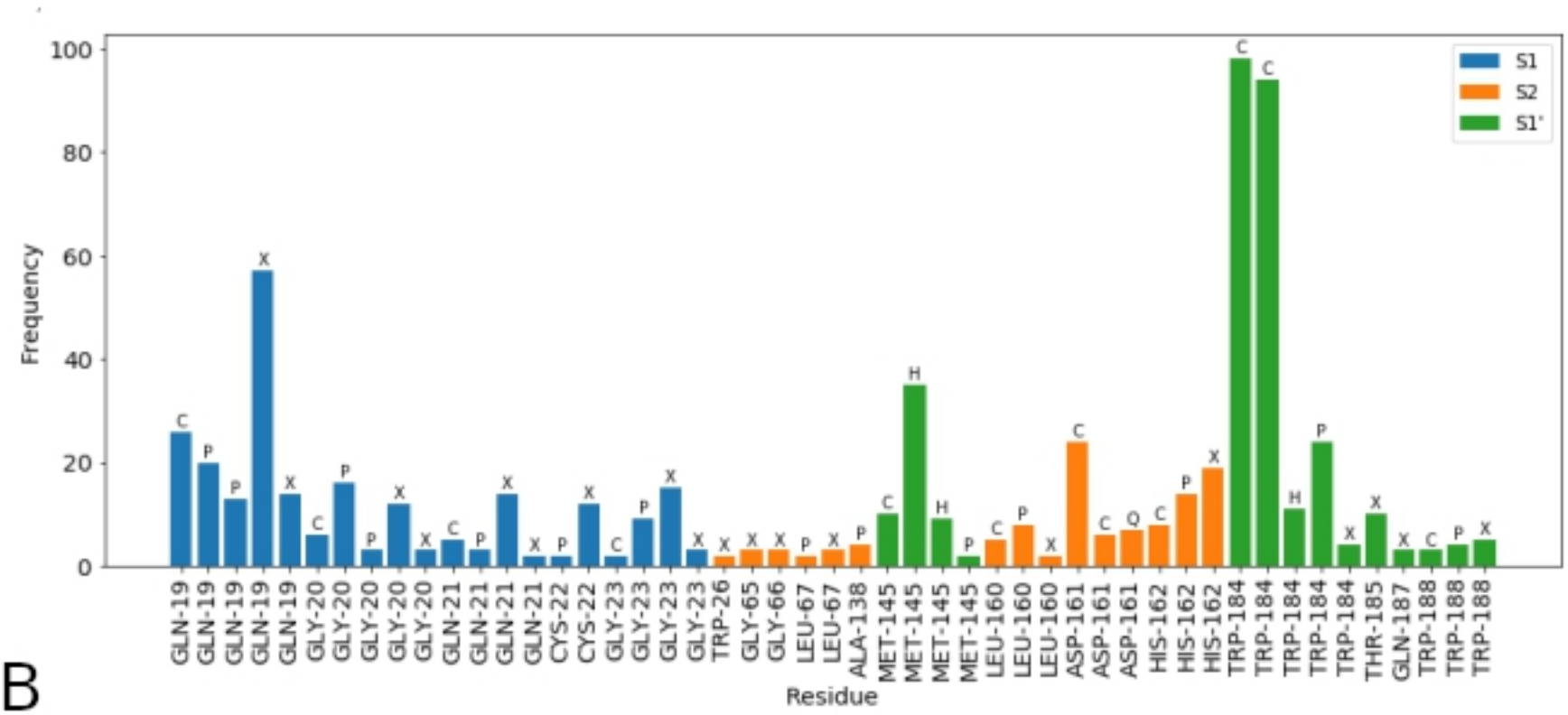

Training Set
Test Set
+2.5SD
−2.5SD
$h_{max} = 0.375$
Standarized Residuals
Leverage Values
3
2
1
0
−1
−2
−3
0.0
0.1
0.2
0.3
0.4
0.5
0.6

figure_5


S1
Catalytic site
S3
S2
S1'
Frequency
Residue
ASP-18
ASP-18
GLN-19
GLN-19
GLY-20
GLY-20
GLY-20
GLY-20
GLN-21
CYS-22
GLY-23
GLY-23
GLY-23
CYS-25
CYS-25
THR-59
THR-59
ASP-60
SER-61
GLY-66
GLY-66
GLY-66
MET-68
ALA-141
MET-145
ASP-161
ASP-161
ASP-161
HIS-162
HIS-162
HIS-162
TRP-184
TRP-184
TRP-184
TRP-184
TRP-184
GLN-187
TRP-188
TRP-188

A


S1
S2
S1'
Frequency
Residue
GLN-19
GLN-19
GLN-19
GLN-19
GLN-19
GLY-20
GLY-20
GLY-20
GLY-20
GLN-21
GLN-21
GLN-21
GLN-21
CYS-22
CYS-22
GLY-23
GLY-23
GLY-23
GLY-23
TRP-26
GLY-65
GLY-66
LEU-67
LEU-67
ALA-138
MET-145
MET-145
MET-145
MET-145
LEU-160
LEU-160
LEU-160
ASP-161
ASP-161
ASP-161
HIS-162
HIS-162
HIS-162
TRP-184
TRP-184
TRP-184
TRP-184
TRP-184
THR-185
GLN-187
TRP-188
TRP-188
TRP-188

B

figure_6

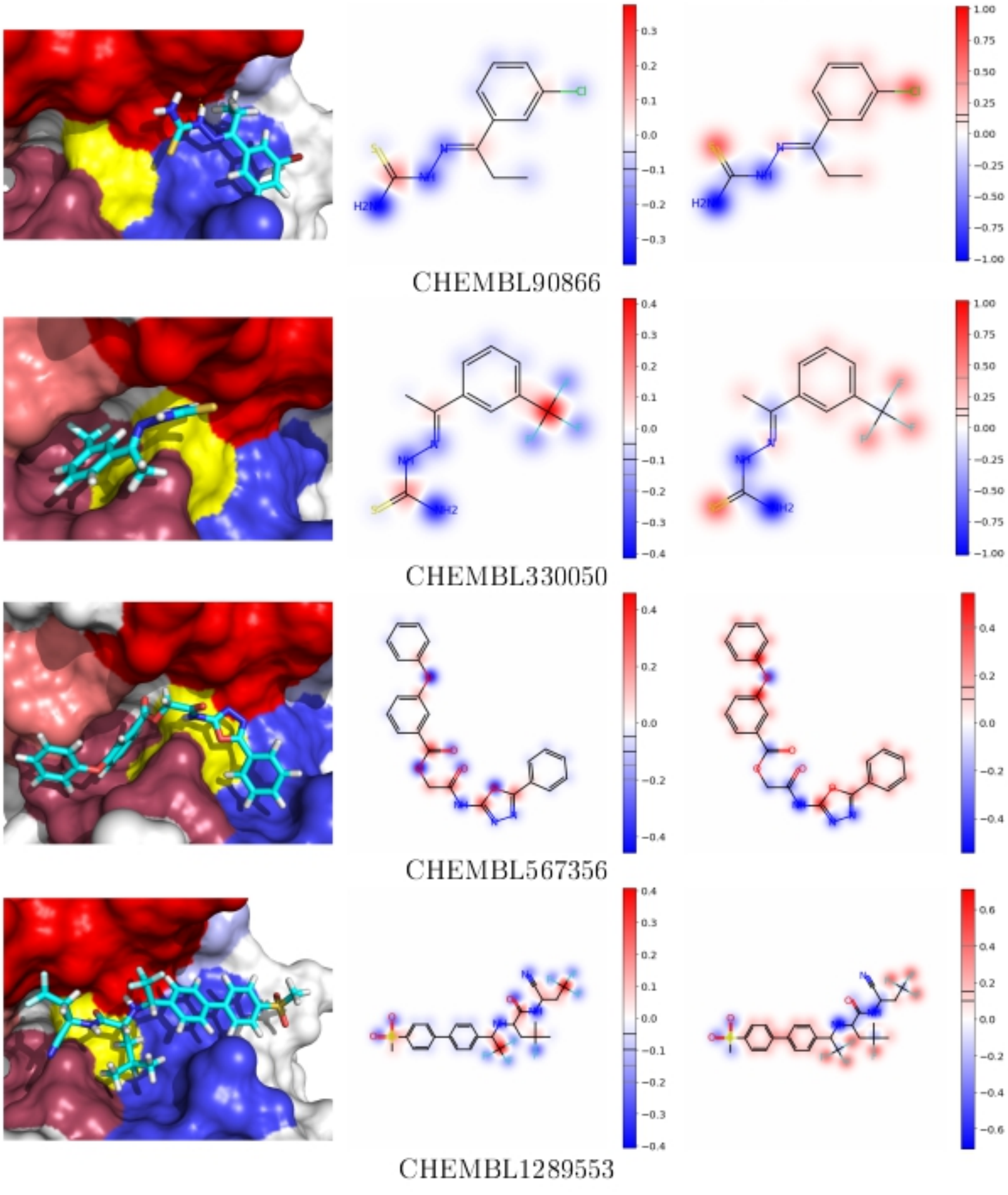
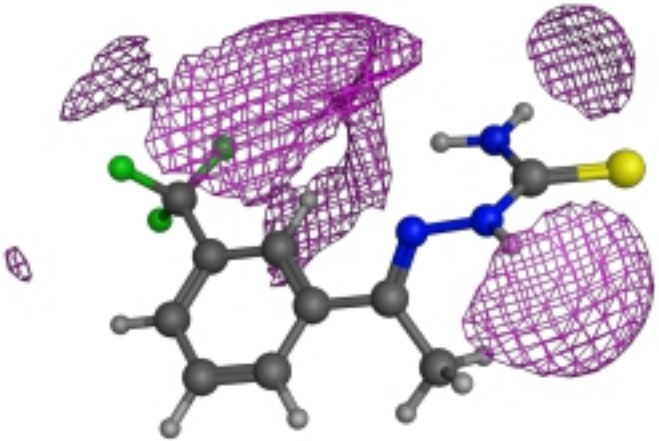
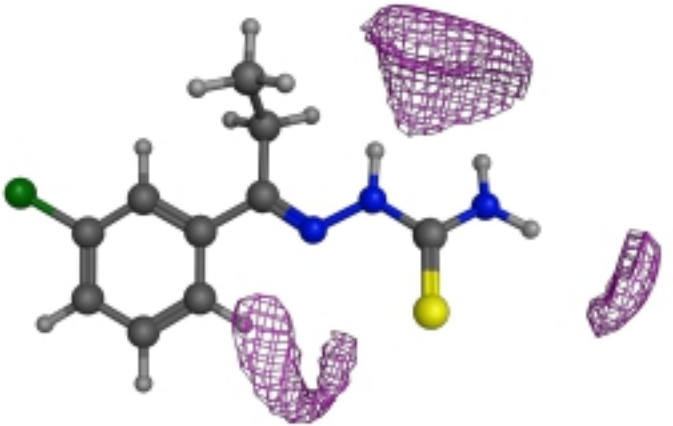


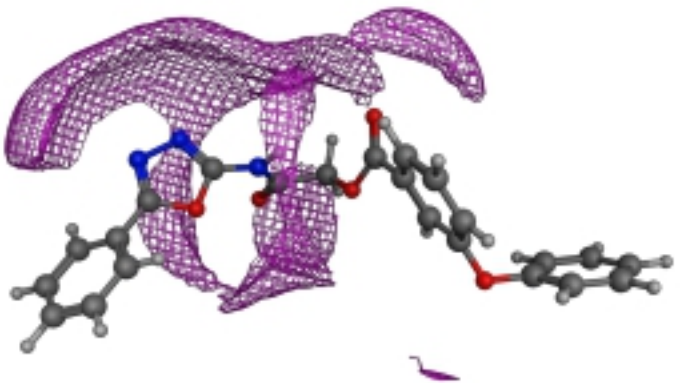
figure_7



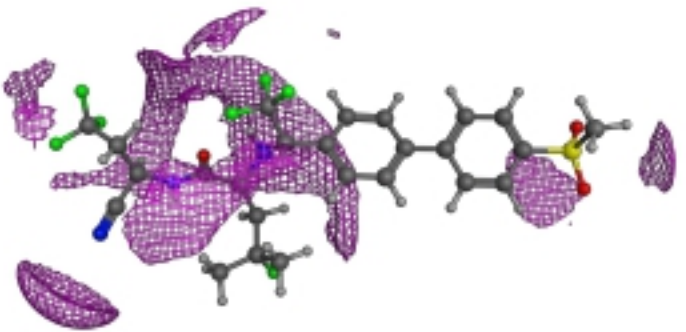
(a) CHEMBL90866



(b) CHEMBL330050



(c) CHEMBL567356



(d) CHEMBL1289553

figure_8