DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Investigating the lack of translation from cruzain inhibition to *Trypanosoma cruzi* activity with machine learning and chemical space analyses

M.Sci. Rafael F. Lameiro,[a] Prof. Dr. Carlos A. Montanari*[a]

Abstract: Chagas disease is a neglected tropical disease caused by the protozoa *Trypanosoma cruzi*. Cruzain, its main cysteine protease, is commonly targeted in drug discovery efforts to find new treatments for this disease. Even though the essentiality of this enzyme for the parasite has been established, many cruzain inhibitors fail as trypanocidal agents. This lack of translation from biochemical to biological assays can involve several factors, including suboptimal physicochemical properties. In this work, we aim to rationalize this phenomenon through chemical space analyses of calculated molecular descriptors. These include statistical tests, visualization of projections, scaffold analysis, and creation of machine learning models coupled with interpretability methods. Our results demonstrate a significant difference between the chemical spaces of cruzain and *T. cruzi* inhibitors, with compounds with more hydrogen bond donors and rotatable bonds being more likely to be good cruzain inhibitors, but less likely to be active on *T. cruzi*. In addition, cruzain inhibitors seem to occupy specific regions of the chemical space that cannot be easily correlated with *T. cruzi* activity, which means that using predictive modeling to determine whether cruzain inhibitors will be trypanocidal is not a straightforward task. We believe that the conclusions from this work might be of interest for future projects that aim to develop novel trypanocidal compounds.

Keywords: chemical space, cruzain, machine learning, *Trypanosoma cruzi*

Introduction

Chagas disease, caused by the protozoa *Trypanosoma cruzi*, is a neglected tropical disease with significant health and social impact on developing countries. The only two drugs available for its treatment, namely, Benznidazole and Nifurtimox, are unsatisfactory due to side effects, toxicity issues and the inability to prevent the progression of symptoms associated with the chronic phase of the disease.[1–3] While there has been progress toward clinical trials for Chagas disease, none of the compounds evaluated so far presented acceptable drug profiles[4]. Therefore, even though phenotypical screening still remains an essential source of potentially bioactive compounds, target-based approaches are being increasingly seen as more promising[4], with the targeting of proteases receiving particular interest, given the successful use of protease inhibitors to treat a wide range of infectious diseases.[5–8]

The most prominent protease inhibitor with *T. cruzi* activity is K777, a highly potent inhibitor of cruzain, its major cysteine protease.[9,10] K777 presented excellent *in vivo* activity in animal models of both acute and chronic Chagas disease, reducing the parasite burden of mice,[11–13] but, in spite of the promising pre-clinical results, it did not progress into clinical trials, primarily due to observed side-effects such as emesis induction and increase in hepatotoxicity biomarkers.[14,15] Currently, the biotechnology company Selva Therapeutics is trying to get K777 (now SLV213) into clinical trials under a new formulation that may resolve the previously observed issues.[4]

The extensive research around K777 has provided a status of "validated target" for cruzain, which has led several research groups to explore cruzain inhibitors as potential new treatments for Chagas disease.[16–20] However, despite the essentiality of cruzain for the *T. cruzi* life cycle, it has been observed in several cases a lack of translation from *in vitro* cruzain activity to *T. cruzi* biological activity (both *in vitro* and *in vivo*).[19,21–24] Some studies have suggested that reversible cruzain inhibitors would be less likely to present trypanocidal activity.[25–27] However, exceptions to this have been presented, such as the two compounds from Merck, Cz007 and Cz008, that are dipeptide-like reversible cruzain inhibitors with excellent trypanocidal activity.[21,28] Nevertheless, not many examples of compounds with satisfactory dual cruzain/*T. cruzi* activity are known. Several factors may be involved in this phenomenon, such as: suboptimal physicochemical properties of the designed compounds, toxicity, the wide diversity of *T. cruzi* phenotypes, or *in vitro* assays that are not good proxies for *in vivo* activity.[4,24,29]

The lack of translation from enzyme to biological activity is a problem of high importance and understanding it would allow for a more rational design of compounds with the desired activity profile against *T. cruzi*. In this context, the use of molecular descriptor analyses and machine learning models can be a good starting point, since these methods have been known to provide valuable insights into several different tasks in medicinal chemistry.[30–33] This is especially true when said models are paired with interpretation techniques that allow for understanding the contribution of each descriptor to the predictions made.[34,35]

[a] *Medicinal and Biological Chemistry Group, São Carlos Institute of Chemistry, University of São Paulo. Trabalhador São-Carlense Avenue, 400, São Carlos, Brazil.*
**e-mail: Carlos.Montanari@usp.br, phone/fax: +55-16-3373-9986/68. Twitter: @CarlosMontanari*

Supporting Information for this article is available on the WWW under www.molinf.com

In this work, we aim to evaluate whether the observed lack of translation from cruzain inhibitors to *T. cruzi* active compounds can be explained as differences in interpretable physicochemical properties and molecular structural motifs. For this, we perform statistical and chemical space analyses of calculated chemical descriptors, analyze chemical scaffolds, and create interpretable classification machine learning models for *T. cruzi* and cruzain active compounds.

Experimental Section

Dataset preparation

The datasets used in this work were downloaded from ChEMBL (version 29)[36] as *csv* files and read into Jupyter Notebooks as Pandas (version 1.3.5) dataframes. For *Trypanosoma cruzi*, only compounds with "Standard Type" equal to IC_{50} were used, while for cruzain, compounds with "Standard Type" equal to Potency, K_i, or IC_{50} were used.
Instances with undefined "Standard Relation" were removed if the value provided did not indicate clearly whether the compound should be classified as active or inactive, according to our criterion (activity < 1 µM for active and > 1 µM for inactive). All activity values were then converted to a unified representation (*p*X) that is the negative of the base-10 logarithm of the activity value in molar units. Prior to conversion to *p*X, K_i values were multiplied by 2.3 to be on a similar scale to IC_{50}, as suggested in the literature,[37] and compounds labeled as Potency were considered as being on the same scale as IC_{50}. Compounds with *p*X out of the usual range (< 3 or > 12) were removed. Then, the SMILES representations of the compounds were standardized using the module chembl_structure_pipeline (version 1.0.0) and canonical SMILES were generated using RDKit (version 2022.03.2).[38] Finally, 200 physicochemical descriptors available in RDKit were calculated.
From this point on, we will be referring to the complete *T. cruzi* and complete cruzain datasets as *tc_complete* and *cz_complete*, respectively. The subsets of compounds with *p*X > 6.0 will be called *tc_active6* and *cz_active6*, while the subsets of compounds with *p*X > 7.0 will be called *tc_active7* and *cz_active7*. Similarly, *_inactive* will be applied to the datasets of compounds with activity below each of the thresholds.

Statistical analyses of descriptor distributions

Distributions of eight interpretable physicochemical descriptors widely used in medicinal chemistry were analyzed. These are the negative base-10 log of the octanol/water partition coefficient (MolLogP), molecular weight (MolWt), topological polar surface area (TPSA), fraction of sp^3 carbon atoms (FCSP3), number of hydrogen bond donors (NumHDonors) and acceptors (NumHAcceptors), number of aromatic rings (NumAromaticRings), and number of rotatable bonds (NumRotatableBonds). All statistical analyses were performed using the module SciPy (version 1.4.1). The functions used are indicated below, in parenthesis.
For continuous descriptors, normality (`normaltest`), skewness (`skew`), and kurtosis (`kurtosis`) were assessed. To determine whether two distributions were sampled from the same population, the Kolmogorov-Smirnov test (`ks_2samp`) was used for continuous descriptors, while for discrete descriptors, the Two-Sample Chi-Squared Test (`chi2_contingency`) was used. For all descriptors, the Mann-Whitney U-Test (`mannwhitneyu`) was used to determine whether samples belonged to the same distributions, although this test depends on the two distributions having the same variance. Therefore, for each pair of datasets compared, equality of the variances for each descriptor was assessed with Levene's test (`levene`). The following pairs of distributions were compared: *tc_complete* x *cz_complete*, *tc_active6* x *tc_inactive6*, *tc_active7* x *tc_inactive7*, *cz_active6* x *cz_inactive6*, *cz_active7* x *cz_inactive7*, *tc_active6* x *cz_active6*, and *tc_active7* x *cz_active7*.
We repeated the tests by removing outliers (compounds outside the interval corresponding to the mean, plus or minus three standard deviations) from both datasets to confirm that distributions were not identified as different due to the presence of outliers.

Chemical space plots

We used the eight main descriptors, as well as all 200, to create chemical space plots for every pair of distributions described in the previous section. After normalizing each column in the dataset by subtracting the mean and dividing by the standard deviation, we used four dimensionality reduction techniques, namely, Principal Components Analysis (PCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) to create and visualize two-dimensional projections of the datasets. The functions `decomposition.PCA`, `manifold.MDS`, and `manifold.TSNE`, from the module scikit-learn (version 1.0.2, hereon referred to as sklearn)[39] were used to calculate PCA, MDS, and t-SNE projections. The module umap (version 0.5.2) was used to calculate UMAP projections. For the projections of active x inactive compounds, a sample of the inactive datasets with the same number of rows as the active datasets was taken instead of using the whole datasets due to the higher number of inactive instances.

Classification models

The following dataset pairs were used as classes for the creation of classification models: *tc_active6* x *tc_inactive6*, *tc_active6* x *cz_active6*, and *tc_active7* x *cz_active*_7. Prior to model fitting, rows that were duplicated for different classes were dropped, collinear descriptors (Pearson's R > 0.7) were checked for (if found, one of the columns was removed), and columns were scaled using sklearn `StandardScaler`.
For all three tasks, compounds were described by the eight main physicochemical descriptors. The following sklearn functions were used to create classification models: `DummyClassifier`, `LogisticRegression`, `SVC`, `RandomForestClassifier`, and `GradientBoostingClassifier`. Model quality was evaluated during training by performing internal validation (sklearn `StratifiedShuffleSplit`) and calculating the following metrics: Accuracy, Area Under ROC Curve, Balanced Accuracy, F1-Score, Matthew's Correlation Coefficient, Precision and Recall. After training, external

validation was performed by calculating the same metrics on a random stratified sample corresponding to 20% of the original datasets.

For the main classification task of this work, *tc_active6* x *cz_active6*, we performed 30 iterations of Bayesian optimization for the model that presented the best metrics using the scikit-optimize (version 0.9.0) function `BayesCV`. Moreover, to understand the effect of training the models on a diverse set of descriptors, feature selection was performed on the 200 RDKit descriptors using sklearn `SelectKBest` (20 descriptors selected), followed by model fitting and calculation of feature importance.

Feature importance was calculated for selected models using the sklearn function `permutation_importance`, with 20 permutations per feature, and using the module SHAP (version 0.40.0).[40]

Proportion of active compounds

To further validate the results from the classification models, we defined usual ranges for each descriptor and evaluated the proportion of active compounds to all compounds tested for bins within these ranges. For instance, the interval (200, 800) was selected for molecular weight, and the proportion of active compounds to all compounds tested was calculated for every 30 units. The proportions were plotted for *T. cruzi* and cruzain active compounds (*p*X > 6.0).

Scaffold analyses

RDKit's implementation of Murcko-type decomposition into scaffolds was used to analyze chemical scaffold distribution for the two targets, specifically, the nature of the scaffolds and the total number of compounds containing the scaffold. The dipeptidyl motif was considered separately, as it is relevant in the context of cruzain inhibitors, but not a Murcko scaffold (not a ring system).

Final validation on a selected external set

To build on observed results, an additional validation step with a selected external set of nine compounds with known activity (not used in training) was performed for the *tc_active6* x *cz_active6* classification model.

Results and Discussion

Statistical analyses of descriptor distributions

After cleaning, standardization, and calculation of chemical descriptors, we found 7867 unique compounds tested for *T. cruzi* activity. Of those, 1793 had an activity greater than 6.0 units on the *p*X scale, and 710 had a *p*X > 7.0. For cruzain, 29333 compounds were tested, with 612 having a *p*X > 6.0 and 240 with a *p*X > 7.0.

Our first data analysis step was to compare the distributions of the eight main chemical descriptors for different pairs of datasets, either considering all compounds tested or only the actives, with *p*X above the two defined thresholds.

At first, Welsh's t-Test, a two-sample comparison test that evaluates whether two populations have the same means, was considered. However, this test assumes that the samples to be compared are normally distributed, which was not the case for most of the samples, some of which also presented significant skewness and kurtosis (see the Supporting Information, available at zenodo.org/record/6876342). Therefore, we restrained our analyses to non-parametric tests.

To compare distributions pairwise, we chose to perform the Mann-Whitney U-Test (U-Test) on all eight descriptors. This is a non-parametric test for which there are no assumptions regarding normality of the distributions but requires continuous or ordinal distributions. In addition, we performed two more tests to compare distributions: the Two-Sample Kolmogorov-Smirnov Test (KS-test), which is suitable for continuous descriptors, and the Two-Sample Chi-Squared Test (X^2-test), for discrete descriptors. For all tests considered, a p-value < 0.05 indicates that there is a statistically significant difference between the compared distributions. Table 1 summarizes the results for all statistical tests, indicating the pair of datasets compared and the descriptors regarded as equal according to each test.

Table 1. Results of statistical tests applied to compare different dataset pairs. The distributions of the indicated descriptors were identified as equal by the corresponding test.

Datasets compared	U-test	KS-test	X^2-test
tc_complete/ cz_complete			
tc_active6/ tc_inactive6		MolLogP FCSP3 TPSA	
cz_active6/ cz_inactive6	TPSA		
tc_active7/ tc_inactive7			
cz_active7/ cz_inactive7	MolLogP* MolWt*		
tc_active6/ cz_active6	MolWt* FCSP3*		
tc_active7/ cz_active7	MolWt*		

*Results from Mann-Whitney U-Test are marked with an asterisk if Levene's test indicated that variables compared have different variances.

When considering all compounds tested on both assays, the sampled chemical space seems to be different for each endpoint, as none of the distributions are regarded as equal. This, however, may just reflect the high number of datapoints within each dataset. Compounds that are active and inactive for the same assay, regardless of how we define them, tend to occupy different regions of the chemical space for most chemical descriptors, as expected. However, a significant overlap in distributions can be observed for most descriptors when we observe the plotted distributions, especially for the *cz_active6* x *cz_inactive6* datasets, since there is an almost 47:1 inactive:active proportion.

Surprisingly, when comparing *T. cruzi* and cruzain inhibitors using both activity thresholds, the tests indicate that most of the distributions are different, with a few being inconclusive (distributions are equal according to the U-test but with different variances). Therefore, this could indicate that, in general, *T. cruzi* and cruzain inhibitors are sampled from different regions of the chemical space. This conclusion, however, is not so clear if we look at the plotted distributions. As an example, we present in Figure 1 the distributions for the descriptor MolLogP for *tc_active6* and *cz_active6*, which the tests indicate as significantly different.

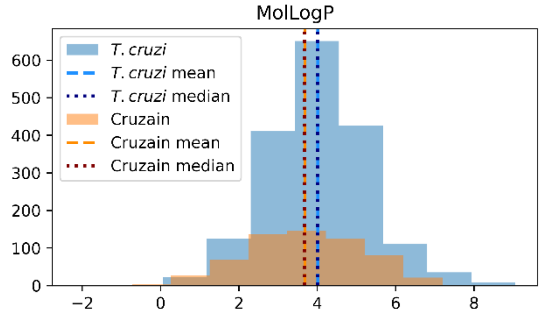


Figure 1. Histograms for the descriptor MolLogP for *T. cruzi* and cruzain inhibitors (*p*X > 6.0).

Even though the distributions are statistically different, it seems that the distribution of cruzain active compounds is within the chemical space of *T. cruzi* inhibitors. We can also observe significant overlaps for the distributions of the other descriptors, as would be expected in a drug discovery context, for which there are commonly used restrictions in the values of descriptors to achieve what has been termed "drug-likeness".[41–43] Therefore, the statistical analyses, coupled with visualization of distributions, are not enough to indicate whether there is a significant difference in the chemical spaces for *T. cruzi* and cruzain inhibitors. The results of the tests and plots, as well as the analyses on the datasets without outliers (for which the conclusions are similar), are provided in the Supporting Information.

Chemical space plots

To evaluate whether differences in the distributions of chemical descriptors would be enough to separate active *T. cruzi* compounds from cruzain inhibitors, we used four different dimensionality reduction techniques to generate two-dimensional projections of the chemical spaces for each pair of datasets compared.

In general, all PCA and MDS projections did not show clear separation of classes or formation of informative patterns and will not be discussed. This is not surprising, considering the significant overlap in the distributions for all of the analyzed chemical descriptors.

The other two methods, t-SNE and UMAP, require the input of a parameter that, in essence, defines whether a projection will give more importance to local or global patterns. For t-SNE, this parameter is called "perplexity", and for UMAP, "n_neighbors". We created plots using five different values for these parameters: 2, 10, 25, 50, and 100. As expected, projections based on values at the lower and higher ends produced graphs with little discernible structure. The values 10, 25, and 50 seemed to produce the most interesting results. In Figure 2, we present t-SNE and UMAP plots for comparing *cz_active6* x *tc_active6* and *cz_active7* x *tc_active7*.

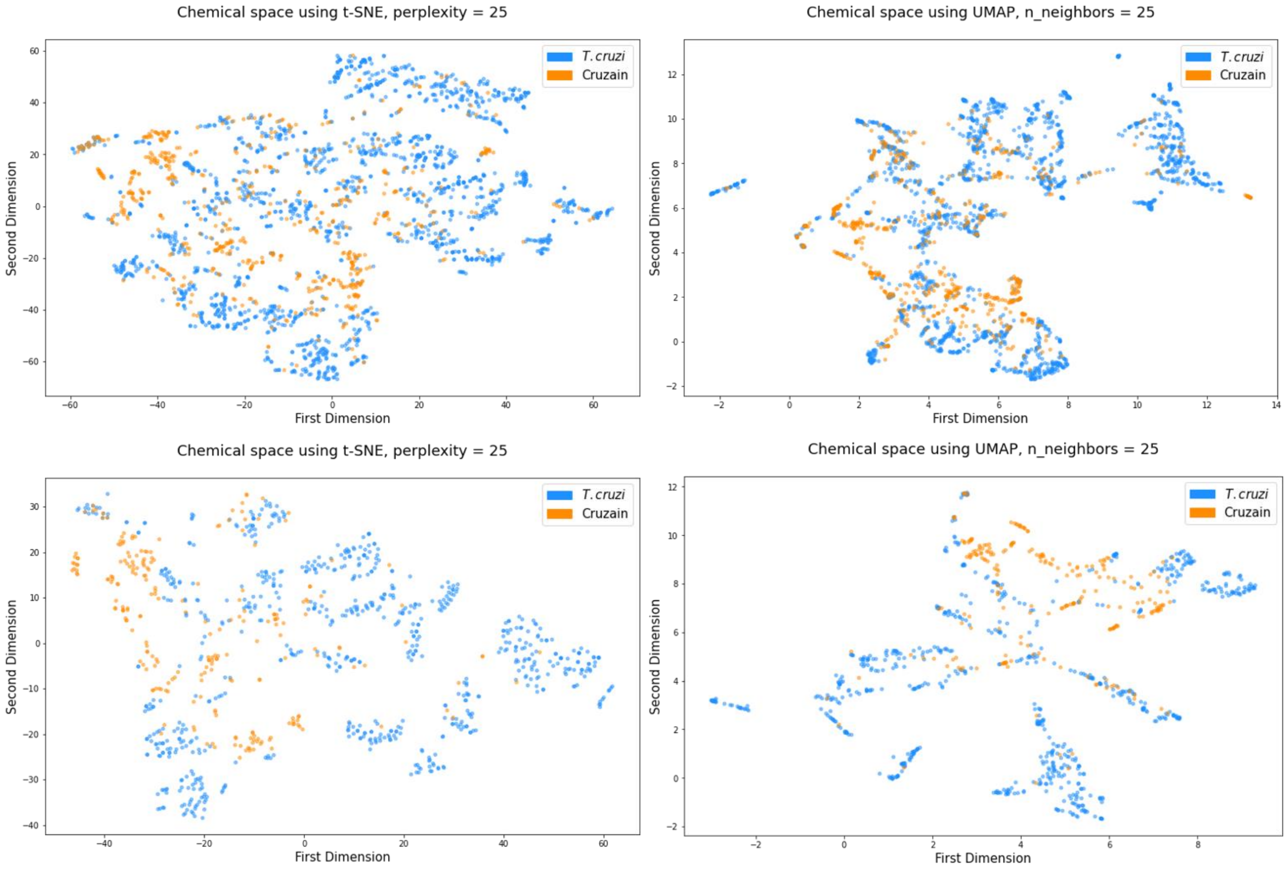


Figure 2. Chemical space plots using t-SNE and UMAP projections (perplexity/n_neighbors = 25) of eight chemical descriptors. Top-left: t-SNE plot for *cz_active6* x *tc_active6*; Top-right: UMAP plot for *cz_active6* x *tc_active6*; Bottom-left: t-SNE plot for *cz_active7* x *tc_active7*; Bottom-right: UMAP plot for *cz_active7* x *tc_active7*.

Before analyzing the plots, it is essential to point out that these dimensionality reduction techniques try to group similar compounds into clusters (local chemical space) and that the distance between clusters may not reflect the distances in the multidimensional space, as they depend on the parameters used for creating the projections. This means that, in general, it is only significant to consider compounds within a cluster as similar. In the four plots shown, we can observe the formation of several clusters of compounds with the same color, both blue and orange. This shows that a significant proportion of the local chemical space of *T. cruzi* inhibitors is not accessed by cruzain inhibitors, suggesting that the combination of chemical descriptors that define good cruzain inhibitors is not generally suitable for *T. cruzi* inhibition. The presence of clusters of cruzain active compounds with no or few *T. cruzi* inhibitors in them also points to the idea that the chemical space of cruzain inhibitors may not always be appropriate to achieve activity against *T. cruzi*. This is even more clear when we only consider the highly active (*p*X > 7.0) compounds. Similar results were observed when all 200 descriptors available on RDKit were used to create the plots. In Figure 3, we show two t-SNE plots with perplexity equal to 25 for *cz_active6* x *tc_active6* and *cz_active7* x *tc_active7*, in which an even more precise separation of cruzain inhibitors on isolated clusters can be perceived.

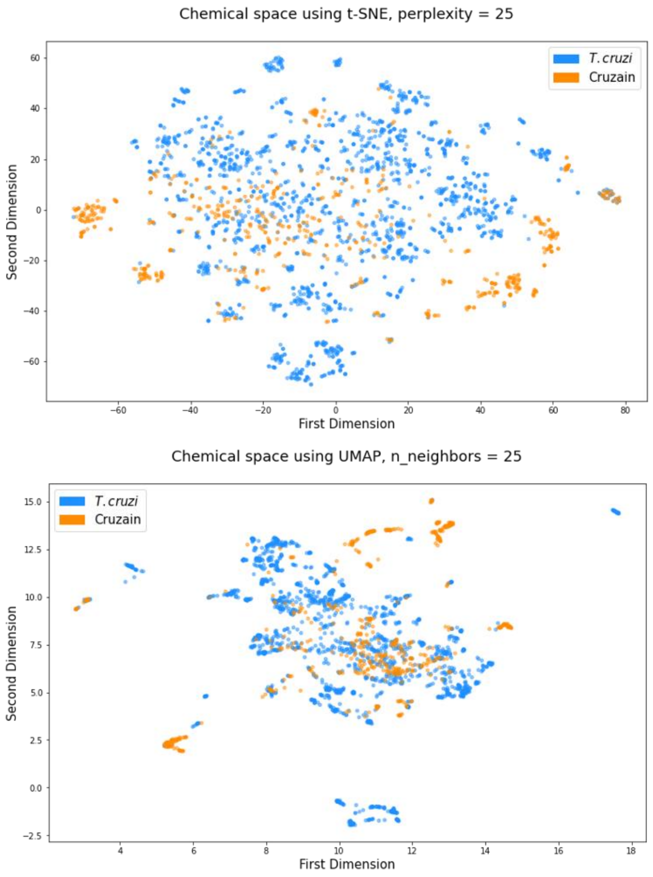


Figure 3. Chemical space plots using t-SNE projections (perplexity = 25) of 200 chemical descriptors available on RDKit. Top: *cz_active6* x *tc_active6*. Bottom: *cz_active7* x *tc_active7*.

We also plotted active (*p*X > 6.0) x inactive compounds for each assay (Figure 4). No clear separation can be seen for compounds tested on *T. cruzi*, perhaps indicating more comprehensive SAR studies for analogs, with several scaffolds having examples of both active and inactive instances. As for cruzain, there seem to exist more clusters of active compounds on the edges of the plots, which could indicate a more restricted chemical space for cruzain inhibition. It is reasonable to assume that the chemical space for inhibitors of a single enzyme will be more restricted (few pharmacophores) than for compounds with biological activity on a parasite, for which several enzymes and biological pathways could be targeted. All plots produced are available in the Supporting Information.

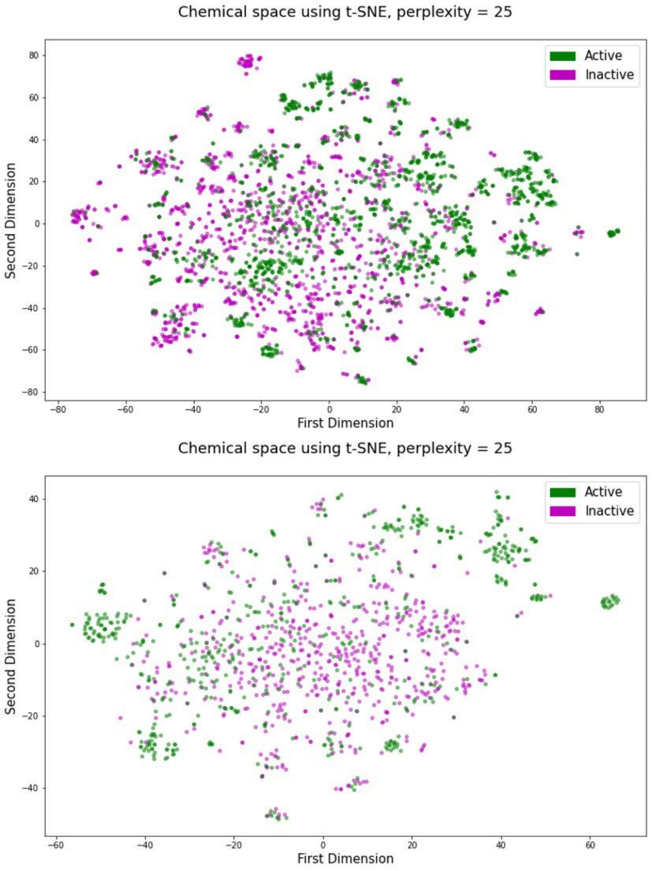


Figure 4. Chemical space plots using t-SNE projections (perplexity = 25) of 200 chemical descriptors for active (*p*X > 6.0) x inactive compounds for *T. cruzi* (top) and cruzain (bottom).

Classification models

Having observed that the chemical spaces of *T. cruzi* active and cruzain active compounds are likely to be different, we moved on to evaluate whether binary machine learning models could efficiently separate the compounds in different classification tasks. We opted to prioritize the eight main chemical descriptors not because they are likely to provide an optimal class separation, but because they are easily interpretable, and to find out if these simpler models could provide enough classification power to allow for the extraction of useful insights.

Model performance was investigated for five machine learning algorithms. Dummy Classifier (DC) is a model that simply classifies all compounds as the majority class, and it is used as a baseline for model performance (any model that extracts valuable information from the data should perform better than the dummy model). The other algorithms are sklearn implementations of Logistic Regression (LR), Support Vector

Machines for classification (SVC), Random Forest (RF), and Gradient Boosting Machines (GBM).

Table 2 shows the calculated metrics (see the legend for the meaning of the abbreviations) for the classification models trained on the eight primary descriptors on three different classification tasks. Two testing conditions were considered: ten rounds of train/test evaluations using sklearn `StratifiedShuffleSplit` (Train) and evaluation on a separate test set not used in training (Test) are reported.

Table 2. Average quality metrics for models trained on the eight main descriptors. The results are for an internal testing with 10 rounds of train/test evaluations using StratifiedShuffleSplit (Train) and for testing on a separate test set not used in training (Test).

	Model	**ACC**	**AUC**	**BACC**	**F1**	**MCC**	**PR**	**REC**
tc_active6 x tc_inactive6	Dummy (Train)	0.77	0.50	0.50	0.00	0.00	0.00	0.00
	Dummy (Test)	0.77	0.50	0.50	0.00	0.00	0.00	0.00
	LR (Train)	0.76	0.70	0.50	0.03	0.00	0.22	0.02
	LR (Test)	0.76	0.50	0.50	0.05	0.00	0.24	0.03
	SVC (Train)	0.81	0.78	0.62	0.40	0.37	0.75	0.27
	SVC (Test)	0.80	0.61	0.61	0.37	0.34	0.69	0.26
	RF (Train)	0.85	0.85	0.72	0.59	0.52	0.75	0.49
	RF (Test)	0.84	0.71	0.71	0.58	0.51	0.74	0.48
	GBM (Train)	0.81	0.80	0.64	0.45	0.38	0.68	0.33
	GBM (Test)	0.82	0.64	0.64	0.45	0.40	0.72	0.33
tc_active6 x cz_active6	Dummy (Train)	0.75	0.50	0.50	0.00	0.00	0.00	0.00
	Dummy (Test)	0.75	0.50	0.50	0.00	0.00	0.00	0.00
	LR (Train)	0.80	0.76	0.65	0.46	0.40	0.75	0.33
	LR (Test)	0.82	0.68	0.68	0.52	0.46	0.77	0.40
	SVC (Train)	0.87	0.87	0.76	0.67	0.62	0.87	0.55
	SVC (Test)	0.86	0.76	0.76	0.66	0.60	0.84	0.55
	RF (Train)	0.87	0.91	0.78	0.69	0.63	0.83	0.60
	RF (Test)	0.88	0.79	0.79	0.72	0.65	0.84	0.62
	GBM (Train)	0.85	0.87	0.75	0.64	0.56	0.78	0.55
	GBM (Test)	0.85	0.75	0.75	0.65	0.58	0.82	0.54
tc_active7 x cz_active_7	Dummy (Train)	0.75	0.50	0.50	0.00	0.00	0.00	0.00
	Dummy (Test)	0.75	0.50	0.50	0.00	0.00	0.00	0.00
	LR (Train)	0.83	0.85	0.72	0.58	0.50	0.73	0.50
	LR (Test)	0.84	0.75	0.75	0.64	0.54	0.71	0.59
	SVC (Train)	0.89	0.93	0.81	0.75	0.69	0.88	0.65
	SVC (Test)	0.86	0.75	0.75	0.66	0.60	0.83	0.54
	RF (Train)	0.91	0.95	0.85	0.80	0.74	0.87	0.74
	RF (Test)	0.88	0.80	0.80	0.73	0.67	0.83	0.65
	GBM (Train)	0.90	0.93	0.84	0.78	0.72	0.86	0.71
	GBM (Test)	0.87	0.79	0.79	0.70	0.62	0.78	0.63

ACC: Accuracy, AUC: Area under curve, BACC: Balanced accuracy, F1: F1-score, MCC: Matthews correlation coefficient, PR: Precision, REC: Recall. The labels used for the models are the same as those presented in the text.

The results show that most models perform better than the baseline DC model, confirming the existence of some chemical space dissimilarity between each of the different classes when the eight main descriptors are considered. Since we have imbalanced classification problems, we will focus on discussing performance by analyzing the balanced accuracy (BACC) metric.

The models' performances show that it is easier to classify compounds as either active on *T. cruzi* or cruzain than it is to predict activity for compounds tested on *T. cruzi*, thus supporting the hypothesis that these compounds occupy significantly different chemical spaces. Moreover, we can see improved or similar scores for models that classify compounds as either *T. cruzi* active or cruzain active when the activity threshold is moved from 6.0 to 7.0, due to the seemingly more restricted chemical spaces for highly active compounds.

Since Random Forest provided the best model for the *tc_active6* x *cz_active6* classification task, we performed Bayes optimization to try to improve its metrics by finding an optimized set of hyperparameters. Surprisingly, the optimized model did not show an improved test set performance, indicating that our original model is already achieving a good generalization. Therefore, further discussions will consider the original Random Forest model trained using sklearn default parameters.

In addition to estimating performance using metrics, we also calculated feature importances for some of the models, an approach that has been gaining popularity in medicinal chemistry works involving machine learning. These methods allow for an interpretation of the models' inner workings and extraction of valuable insights.[35,44,45] Permutation importance is a method that quantifies the impact of each feature on a final model's performance by shuffling the values of each descriptor column, refitting the model, and evaluating the drop on a chosen quality metric. This can be performed on the same dataset used for training, although it is also recommended that a separate test set is used since high feature importance on the training set that does not show equal importance on the test set may indicate an overfitted model. We also employed SHAP (SHapley Additive exPlanations), a method for quantifying model importance derived from game theory that has been explored in medicinal chemistry projects.[46,47] In summary, a baseline prediction is computed for all datapoints, and the numerical value of every prediction can be distributed as additions or subtractions to the baseline, proportionally to the contribution of each descriptor. Therefore, descriptors that get larger SHAP values impact more the value of a prediction and are regarded as more important. Both global and local interpretations are available for this method.

Instead of calculating feature importance for a single model, we opted to average the results for 50 models trained on different train/test splits. This was done to reduce the impact (or bias) an individual split can have on the model. We present in Figure 5 the bar plots of the average permutation feature importance and global SHAP values for the 50 Random Forest *tc_active6* x *cz_active6* models, along with bar plots for four descriptors that received higher feature importance scores: NumHDonors, NumRotatableBonds, MolWt, and NumAromaticRings.

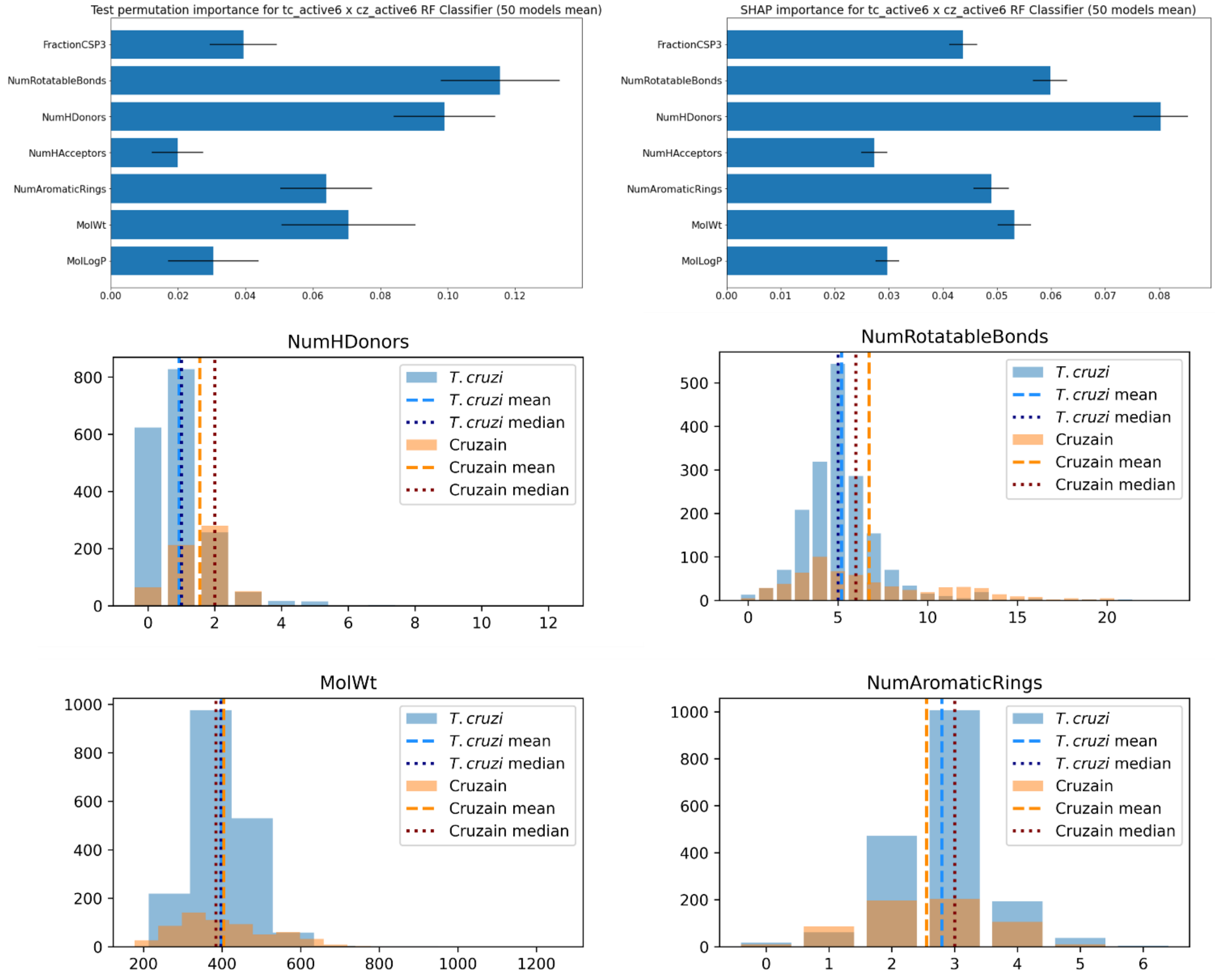


Figure 5. Bar plots of average (with standard deviation) permutation feature importance and global SHAP values for 50 Random Forest *tc_active6* x *cz_active6* models and bar plots for four of the most important descriptors: NumHDonors, NumRotatableBonds, MolWt, and NumAromaticRings.

Overall, both methods agree on finding greater feature importances for the descriptors NumHDonors and NumRotatableBonds, followed by MolWt and NumAromaticRings. These results, combined with an analysis of the bar plots, suggest that compounds that are active on *T. cruzi* tend to have more aromatic rings and a lower number of hydrogen bond donors and rotatable bonds compared to compounds active on cruzain. There also seems to be a higher proportion of compounds with MolWt < 400 among the *T. cruzi* actives. As for the other four descriptors, MolLogP, NumHAcceptors, and FCSP3 tend to show a smaller importance score, as long as they are within the ranges of the training set, while TPSA was removed from model building as it was correlated with NumHAcceptors.

These observations can be rationalized in part if we analyze the active site of cruzain. Figure 6 shows the interaction of a peptide-like compound with the active site of cruzain. Hydrogen bonds are indicated as magenta lines, and the three hydrogen bond donor groups in the inhibitor are marked with magenta arrows.

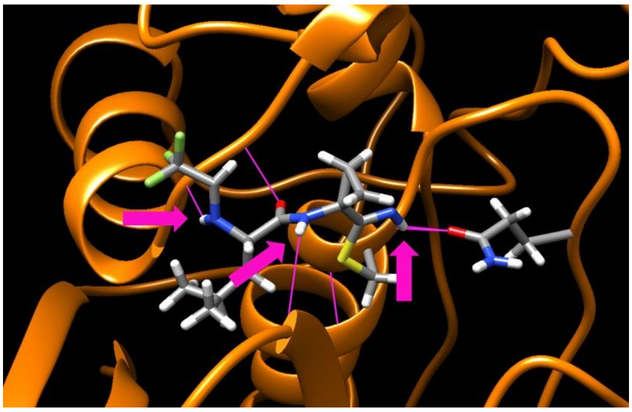


Figure 6. Cruzain inhibitor Neq0981 interacts with hydrogen bond acceptor groups within the active site. Hydrogen bonds are shown as magenta lines and magenta arrows indicate hydrogen bond donor groups in the inhibitor. Parts of the inhibitor were removed for better visualization. Adapted from Lameiro et al.[48]

We can see that the three hydrogen bond donor groups of the peptide-like inhibitor interact with hydrogen bond acceptors on the enzyme (some are backbone groups and are not visible in the image). These favorable interactions are commonly exploited in the development of cruzain inhibitors, mostly to achieve specificity,[49] and the bar plot for NumHDonors shows that most active cruzain inhibitors have one or two hydrogen bond donor groups, while most *T. cruzi* inhibitors have zero or one. Besides, it is common to decorate the scaffold of a cruzain inhibitor with groups that approach the most relevant subsites: P1', P1, P2, and P3, which usually makes use of flexible linking groups that allow for a better adjustment of the compound within the active site. Even though this works for cruzain inhibition, it negatively impacts *T. cruzi* activity, which seems to require more rigid compounds, as the bar plots for NumRotatableBonds show. The number of rotatable bonds is known to impact oral bioavailability, with a classical study having shown that 60% of the evaluated compounds with less than 7 rotatable bonds presented an improved oral availability profile, whereas only 20-30% of the more flexible compounds were classified as such.[42] Therefore, designing more rigid cruzain inhibitors may not only improve the chances of achieving *T. cruzi* activity, but may also improve the chances of a compound being orally available, a fundamental feature of new treatments for Chagas disease, as proposed on DNDi's Target Product Profile.[50,51] Also of interest is the lower importance of the number of hydrogen acceptors and MolLogP. No clear trend can be discerned from the bar plot/histogram other than that the two classes seem to have approximately the same average NumHAcceptors, around five. MolLogP seems to be more important for the higher activity threshold of 7.0, with compounds highly active for *T. cruzi* having a greater mean MolLogP (4.5) than cruzain inhibitors (3.9).

We opted to focus on models based on the eight primary descriptors not only because they are easily interpretable, but also because using all 200 RDKit descriptors would increase the likelihood of spurious correlations appearing. Nevertheless, we wanted to evaluate the performance our models would present if any descriptor combination was available, regardless of interpretability. To avoid the risk of implementing overfitted models, a feature selection approach was used to identify the most important descriptors for the *tc_active6* x *cz_active6* classification task. After removing null-variance and highly collinear descriptors, we used sklearn `SelectKBest` to select the 20 most important descriptors for the classification task and fitted five models as described above.

As expected, some of the variables selected by this approach are highly specific, such as the number of heterocycles, presence of specific fragments, and less interpretable descriptors like BCUT and VSA. The BCUT metrics are a combination of molecular descriptors that take into account atom types, proximity measures, Gasteiger charges, and molecular refractivity, while VSA (Van der Waals Surface Area) descriptors are electrotopological and consider sums of atomic properties (such as Gasteiger charges for PEOE_VSA) within a specific range.[52,53] Being whole-molecule, multiparameter descriptors, they can be especially useful for tasks such as toxicity prediction and quantification of molecular diversity, but challenging to use as a parameter in a lead optimization context.

The best model fit using these descriptors presented a balanced accuracy score of 0.85 on the test set, a 0.06 increase compared to the RF model trained on the eight main descriptors, showing that the interpretation of the models based on the eight descriptors models can be used to get insights about the chemical space variability within *T. cruzi* and cruzain inhibitors, as not much information is added when we include less interpretable descriptors. The results of this analysis can be found in the Supporting Information.

Proportion of active compounds

To further investigate the trends suggested by the previous results, we tried to understand how the number of active compounds varied as each descriptor value progressively increased. For this, the proportion of active compounds (pX > 6.0) over all compounds tested for different ranges of descriptor values was calculated.

Confirming previously observed results, Figure 7 shows that the maximum proportion of active *T. cruzi* compounds occurs at lower values of NumAromaticRings. For NumRotatableBonds, we can see that the *T. cruzi* and cruzain plots intersect, which reinforces the hypothesis that more rigid compounds are preferable for *T. cruzi* activity. The NumHDonors plot for *T. cruzi* indicates a decreasing trend but has local maxima for 5 and 7 hydrogen bond donors, which are likely artifacts due to low sampling (few compounds were tested with these values of NumHDonors, and they happened to be active). No clear trend is shown by NumHAcceptors, other than a maximum at the value 7, which might also indicate a sampling artifact.

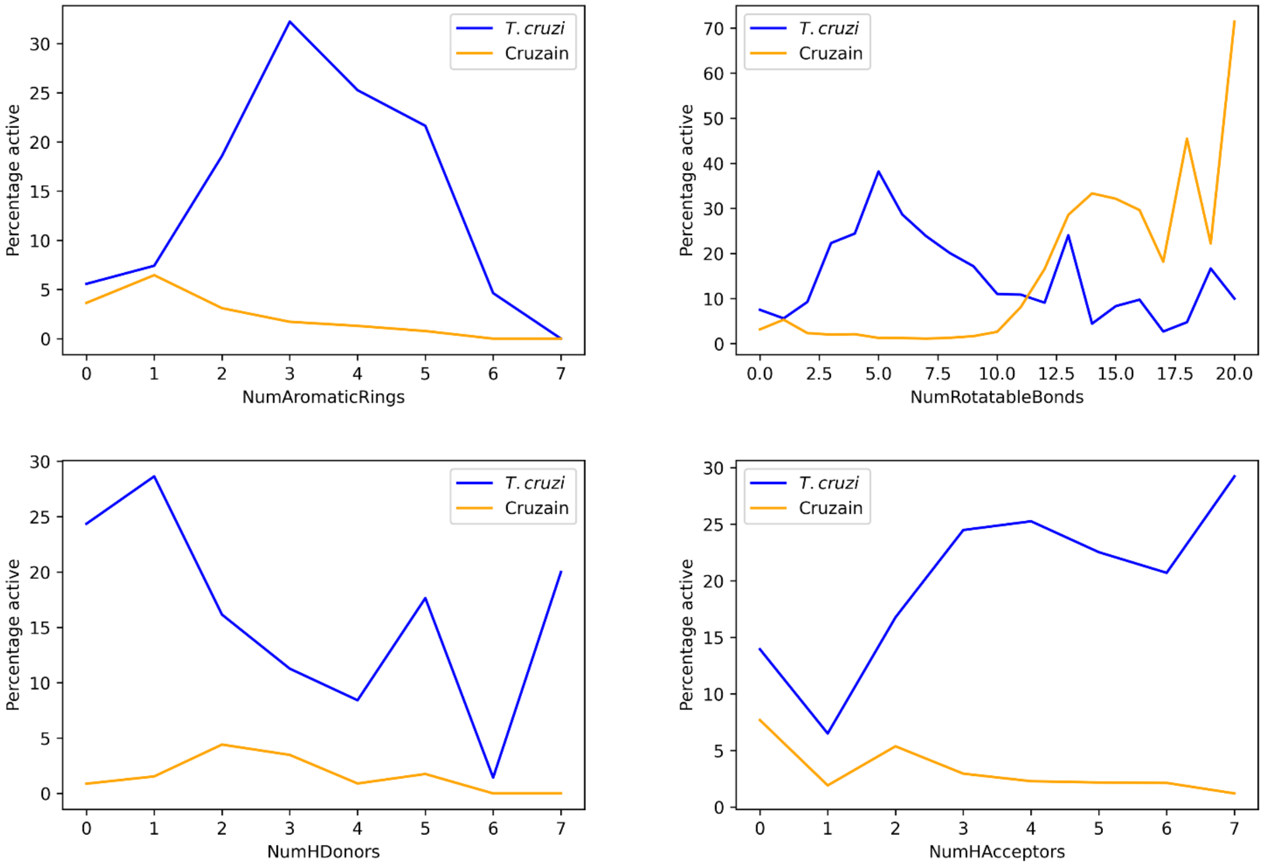


Figure 7. Percentage of active compounds over all tested compounds for different values of discrete descriptors.

Scaffold analyses

A chemical scaffold can be defined as the "structural core of a chemical compound".[54] In medicinal chemistry, when an active chemical compound is identified, it is common to perform minor modifications to its chemical structure while preserving the "core", since its structural and geometrical properties are likely responsible for the observed activity.

RDKit provides an implementation of the Bemis-Murcko method, which depletes molecular structures of their side chains, keeping only the ring(s) framework as a central core or scaffold.[55] This algorithm was used to calculate the total number of scaffolds for the complete datasets and for the subset of active compounds.

The chemical space of the compounds tested on cruzain was diverse, with 10096 out of 29333 compounds having a unique scaffold. For *T. cruzi*, the proportion was similar, with 2640 out of 7867 compounds having a unique scaffold. However, the absolute numbers indicate that the explored chemical space for cruzain inhibitors was more comprehensive than for *T. cruzi*.

We also analyzed the top 10 scaffolds with the greatest number of examples for the active and highly active datasets, evaluating the proportion of active compounds over all compounds per scaffold, to understand whether scaffolds that correspond to more active compounds really contain privileged substructures or if they just reflect compounds that were tested more frequently. The figures generated are provided in the Supporting Information.

Agreeing to our previous results, several of the scaffolds for compounds that are active on *T. cruzi* have three or more aromatic rings and few hydrogen bond donor groups. Most scaffolds have a high proportion of active compounds and seem to represent truly privileged substructures, as shown on Figure 8. For cruzain, several scaffolds contain peptidic bonds or two hydrogen bond donor groups. In addition, we can see that the benzene scaffold is the most frequent for cruzain active compounds, even though the percentage of actives over all tested compounds is low, suggesting that the high number of actives may just reflect a higher number of tested compounds with that scaffold.

We also observed that the benzene scaffold represents not only low molecular weight aromatic structures, but also long-chain peptide-like structures containing one aromatic ring. For these, the aromatic ring likely matches the S2 or S3 subsite of cruzain, where it can interact with hydrophobic side chains. Being cruzain a protease, it is expected that peptidic or peptide-like structures are good fits for its active size. However, since chains of peptidic bonds are not conventional Bemis-Murcko scaffolds (as they do not contain any ring systems), we performed a separate analysis of the dipeptidyl scaffold for cruzain, as well as for *T. cruzi*, using RDKit's `GetSubstructMatches` functionality.

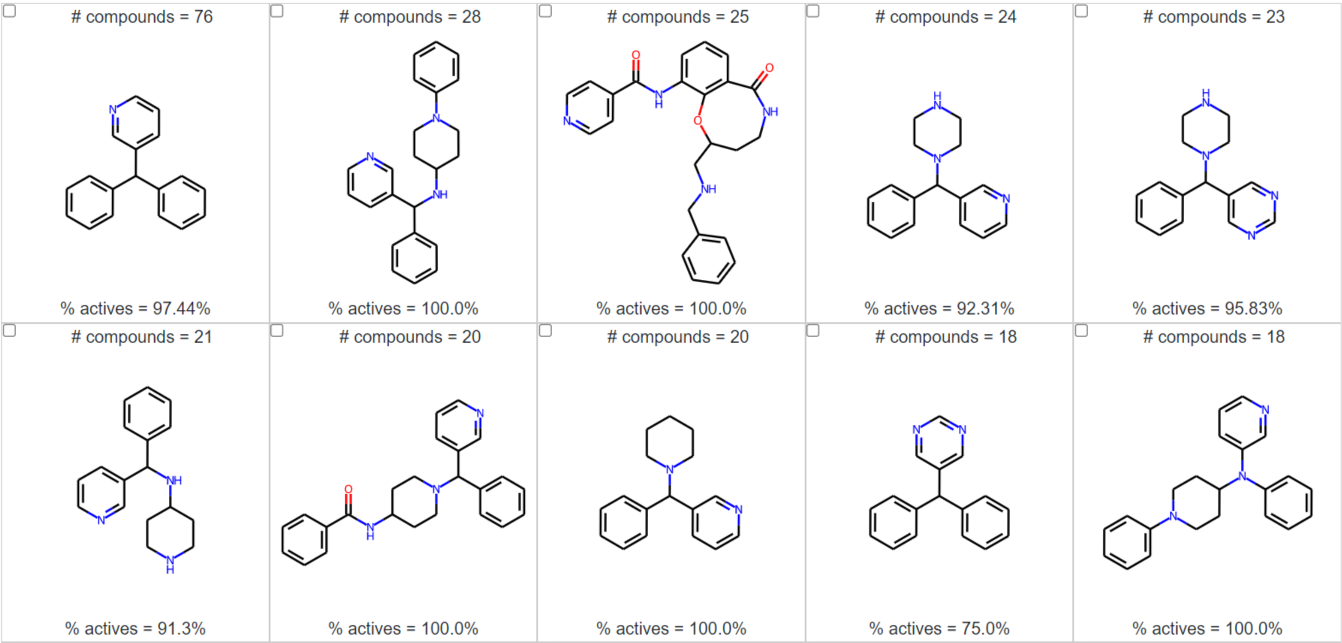


Figure 8. Most populated scaffolds with *T. cruzi* activity > 6.0. Text above: total number of compounds containing that scaffold; Text below: Percentage of active compounds over all compounds tested for that scaffold.

In Table 3, we report the total number of compounds tested, as well as active ($pX > 6.0$) and highly active ($pX > 7.0$) compounds that match this non-conventional scaffold.

Table 3. Analysis of compounds containing the dipeptidyl scaffold.

	Cruzain	***T. cruzi***
Tested compounds with dipeptidyl scaffold	1640 (5.59%)	86 (1.09%)
Number of active compounds ($pX > 6.0$)	169 (10.34%)	6 (6.98%)
Number of highly active compounds ($pX > 7.0$)	104 (6.34%)	1 (1.16%)

The number of dipeptidyl compounds tested on cruzain is, in proportion to the total number of compounds tested, five times higher than for *T. cruzi*. As expected, there is a high proportion of active compounds containing this scaffold for cruzain, but, interestingly, very few compounds are active on *T. cruzi*. In fact, only 1 out of 86 dipeptidic compounds tested on *T. cruzi* were highly active, and none of the *T. cruzi* top 20 scaffolds contained a dipeptidyl motif. Even if we admit that the analysis is not entirely fair, since only two of the highly active cruzain inhibitors (Figure 9) had their activity measured on *T. cruzi* (and therefore, we cannot know how many of them would be active against *T. cruzi*), the data seems to show a trend of low *T. cruzi* activity for peptide-like compounds. This corroborates the analyses that indicated that compounds with more hydrogen bond donors and rotatable bonds, as is the case for most peptidic compounds, are less likely to be active on *T. cruzi*, as well as previous results from our research group.

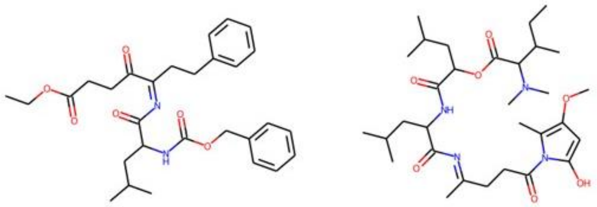


Figure 9. Compounds matching the dipeptidyl scaffold that were highly active on cruzain ($pX > 7.0$) but inactive on *T. cruzi*.

Final validation on a selected external set

To validate the insight derived from the previous results, we analyzed the structures of nine compounds with known *T. cruzi* and cruzain activity profiles. For this, we refitted our final Random Forest *tc_active6* x *cz_active6* model with all compounds available, except those present in this final validation set, to prevent data leakage. The model outputs a probability value (P) between 0 and 1, with values closer to 0 indicating *T. cruzi* active compounds and close to 1, cruzain active compounds. Usually, a threshold of 0.5 is used to create a binary output for the classification model. Adjusting this threshold is possible, but this leads to a higher percentage of false positives or true negatives.

Out of the six compounds with known *T. cruzi* activity analyzed, only Fexinidazole was incorrectly predicted as a cruzain inhibitor ($P = 0.66$). Benznidazole, Nifurtimox, Posaconazole, Ravuconazole, and Neq720[56] are all classified as *T. cruzi* active ($P < 0.5$). Therefore, the model can reasonably identify *T. cruzi* active compounds. The complete table of compounds analyzed can be found in the Supporting Information.

As for K777 (SLV213), Cz007, and CHEMBL4442543,[57] we could expect that our model would predict probabilities close to 0.5, as the compounds would have characteristics of both cruzain and *T. cruzi* active compounds. This was not observed, and all compounds are classified as cruzain inhibitors with high probability, which means that they are closer to other cruzain inhibitors in the chemical space, and not to known *T. cruzi* inhibitors.

This trend of cruzain inhibitors being more similar to other cruzain inhibitors irrespective of the *T. cruzi* activity suggests that using predictive modeling to determine whether cruzain inhibitors will be trypanocidal is not a straightforward task, at least when considering the eight main descriptors evaluated in this work. In other words, the model will likely not be able to distinguish whether cruzain inhibitors will be active or inactive on *T. cruzi*. This is an important observation, with two possible conclusions. One possibility is that cruzain inhibitors occupy specific regions of the chemical space that cannot be correlated with *T. cruzi* activity. If that is the case, then using machine learning with interpretable models to develop better performing cruzain inhibitors is not a feasible task. On the other hand, it is possible that cruzain is not an appropriate target for *T. cruzi* activity and the rare compounds that present bioactivity could be inhibiting other targets or pathways. This would explain, for instance, why more reactive covalent inhibitors usually show more activity than reversible covalent inhibitors.

One final question that naturally arises from all the analyses performed is: would it be possible to modify the structures of cruzain inhibitors to make them more like *T. cruzi* inhibitors? This could be answered by looking at the confusion matrix from our *tc_active6* x *cz_active 6* Random Forest model. Around 12% of the test set compounds are mislabeled. This clearly shows that there is a reasonable overlap of the chemical spaces for the two classes, although this does not necessarily mean that every cruzain inhibitor can be optimized for *T. cruzi* activity; some scaffolds might not respond well to structural modifications without losing cruzain activity (which is the proxy endpoint for *T. cruzi* activity on lead optimization projects). On the other hand, the prevalence of correct classifications should at least serve as a warning to researchers that aim to develop *T. cruzi* active compounds and plan on using cruzain as a target, as the chemical space accessed by the new series may not correspond to an appropriate chemical space for bioactivity against the parasite.

Conclusion

Understanding the lack of translation from activity in cruzain to activity against *T. cruzi* is a highly relevant task for the development of novel treatments for Chagas disease, but it is certainly not a trivial one. Our statistical and machine learning-based analyses show that cruzain inhibitors present distributions of the eight primary chemical descriptors for medicinal chemistry that are, in significant part, within the expected values for drug-like compounds, as well as for compounds with trypanocidal activity. However, chemical space plots and classification models indicate that cruzain inhibitors may, in significant part, populate isolated areas of the chemical space relative to *T. cruzi* inhibitors. There seems to exist an especially relevant mismatch for the values of three chemical descriptors, which are identified as essential to differentiate *T. cruzi* active compounds from cruzain active compounds. As our results show, compounds with high *T. cruzi* activity usually present a higher number of aromatic rings and a lower number of hydrogen bond donors and rotatable bonds. Several compounds with these characteristics represent privileged substructures for *T. cruzi* activity, as indicated by the chemical scaffolds analysis.

In summary, our results suggest that a significant part of the chemical space associated with cruzain activity does not intersect with the chemical space of known *T. cruzi* active compounds, which helps to explain the lack of translation of *in vitro* activity for inhibition of the enzyme to *in vitro*/*in vivo* trypanocidal activity. These are worrying results, considering that the end goal of every project that targets cruzain is the development of a new trypanocidal lead compound or drug. Our analyses also serve as a warning for future researchers working on projects targeting cruzain, as it would be advisable to confirm whether the compounds proposed are within with the chemical space of *T. cruzi* active compounds.

Supporting Information

All datasets used in this study are provided as csv/tsv files. Jupyter Notebooks containing the code necessary for reproducing the data curation processes and analyses are also provided. This material was uploaded to the Zenodo repository (zenodo.org/record/6876342).

Acknowledgments

R.F.L. receives a PhD Scholarship from the São Paulo Research Foundation (FAPESP) - grant #2021/01633-3. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. FAPESP grant #2013/18009-4 is also acknowledged.

References

[1] J. A. Urbina, *Acta Trop.* **2010**, *115*, 55–68.
[2] M. Boiani, L. Piacenza, P. Hernández, L. Boiani, H. Cerecetto, M. González, A. Denicola, *Biochem. Pharmacol.* **2010**, *79*, 1736–1745.
[3] C. N. Paiva, E. Medei, M. T. Bozza, *PLoS Pathog.* **2018**, *14*, 1–19.
[4] A. B. Vermelho, G. C. Rodrigues, C. T. Supuran, *Expert Opin. Drug Discov.* **2020**, *15*, 145–158.
[5] E. C.Y. Toh, N. L. Huq, S. G. Dashper, E. C. Reynolds, *Curr. Protein Pept. Sci.* **2011**, *11*, 725–743.
[6] A. A. Agbowuro, W. M. Huston, A. B. Gamble, J. D. A. Tyndall, *Med. Res. Rev.* **2018**, *38*, 1295–1331.
[7] J. H. McKerrow, *PLoS Negl. Trop. Dis.* **2018**, *12*, 1–3.
[8] L. Cianni, C. W. Feldmann, E. Gilberg, M. Gütschow, L. Juliano, A. Leitão, J. Bajorath, C. A. Montanari, *J. Med. Chem.* **2019**, *62*, 10497–10525.
[9] V. Duschak, A. Couto, *Curr. Med. Chem.* **2009**, *16*, 3174–3202.
[10] L. G. Ferreira, A. D. Andricopulo, *Pharmacol. Ther.* **2017**, *180*, 49–61.
[11] P. S. Doyle, Y. M. Zhou, J. C. Engel, J. H. McKerrow, *Antimicrob. Agents Chemother.* **2007**, *51*, 3932–3939.
[12] J. H. McKerrow, J. C. Engel, C. R. Caffrey, *Bioorg.*

Med. Chem. **1999**, *7*, 639–644.

[13] J. C. Engel, P. S. Doyle, I. Hsieh, J. H. McKerrow, *J. Exp. Med.* **1998**, *188*, 725–734.

[14] J. H. McKerrow, *PLoS Negl. Trop. Dis.* **2018**, *12*, e0005850–e0005850.

[15] J. McKerrow, P. Doyle, J. Engel, L. Podust, S. Robertson, R. Ferreira, T. Saxton, M. Arkin, I. Kerr, L. Brinen, C. Craik, *Mem. Inst. Oswaldo Cruz* **2009**, *104*, 263–269.

[16] H. J. Wiggers, J. R. Rocha, W. B. Fernandes, R. Sesti-Costa, Z. A. Carneiro, J. Cheleski, A. B. F. da Silva, L. Juliano, M. H. S. Cezari, J. S. Silva, J. H. McKerrow, C. A. Montanari, *PLoS Negl. Trop. Dis.* **2013**, *7*, 1–11.

[17] M. Siklos, M. BenAissa, G. R. J. Thatcher, *Acta Pharm. Sin. B* **2015**, *5*, 506–519.

[18] D. A. Nicoll-Griffith, *Expert Opin. Drug Discov.* **2012**, *7*, 353–366.

[19] A. C. B. Burtoloso, S. de Albuquerque, M. Furber, J. C. Gomes, C. Gonçalez, P. W. Kenny, A. Leitão, C. A. Montanari, J. C. Quilles, J. F. R. Ribeiro, J. R. Rocha, *PLoS Negl. Trop. Dis.* **2017**, *11*, 1–16.

[20] L. Cianni, C. Lemke, E. Gilberg, C. Feldmann, F. Rosini, F. D. R. Rocho, J. F. R. Ribeiro, D. Y. Tezuka, C. D. Lopes, S. de Albuquerque, J. Bajorath, S. Laufer, A. Leitão, M. Gütschow, C. A. Montanariid, *PLoS Negl. Trop. Dis.* **2020**, *14*, 760736.

[21] M. Ndao, C. Beaulieu, W. C. Black, E. Isabel, F. Vasquez-Camargo, M. Nath-Chowdhury, F. Massé, C. Mellon, N. Methot, D. A. Nicoll-Griffith, *Antimicrob. Agents Chemother.* **2014**, *58*, 1167–1178.

[22] L. A. A. Avelar, C. D. Camilo, S. De Albuquerque, W. B. Fernandes, C. Gonçalez, P. W. Kenny, A. Leitão, J. H. McKerrow, C. A. Montanari, E. V. M. Orozco, J. F. R. Ribeiro, J. R. Rocha, F. Rosini, M. E. Saidel, *PLoS Negl. Trop. Dis.* **2015**, *9*, 1–24.

[23] F. Altamura, R. Rajesh, C. M. C. Catta-Preta, N. S. Moretti, I. Cestari, *Drug Dev. Res.* **2020**, *n/a*, DOI 10.1002/ddr.21664.

[24] J. M. Kratz, K. R. Gonçalves, L. M. D. Romera, C. B. Moraes, P. Bittencourt-Cunha, S. Schenkman, E. Chatelain, S. Sosa-Estani, *Mem. Inst. Oswaldo Cruz,* **2021**, *116*, 1–11.

[25] D. Luci, W. Lea, R. Ferreira, B. Shoichet, A. Simeonov, A. Rodriguez, A. Jadhav, D. J. Maloney, in *Probe Reports from NIH Mol. Libr. Progr.*, Bethesda (MD), **2010**.

[26] D. G. Silva, J. F. R. Ribeiro, D. De Vita, L. Cianni, C. H. Franco, L. H. Freitas-Junior, C. B. Moraes, J. R. Rocha, A. C. B. Burtoloso, P. W. Kenny, A. Leitão, C. A. Montanari, *Bioorg. Med. Chem. Lett.* **2017**, *27*, 5031–5035.

[27] V. Bonatto, P. H. J. Batista, L. Cianni, D. De Vita, D. G. Silva, R. Cedron, D. Y. Tezuka, S. De Albuquerque, C. B. Moraes, C. H. Franco, J. Lameira, A. Leitão, C. A. Montanari, *RSC Med. Chem.* **2020**, *11*, 1275–1284.

[28] C. Beaulieu, E. Isabel, A. Fortier, F. Massé, C. Mellon, N. Méthot, M. Ndao, D. Nicoll-Griffith, D. Lee, H. Park, W. C. Black, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 7444–7449.

[29] K. Brak, P. S. Doyle, J. H. McKerrow, J. A. Ellman, *J. Am. Chem. Soc.* **2008**, *130*, 6404–6410.

[30] M. T. Kim, A. Sedykh, S. K. Chakravarti, R. D. Saiakhov, H. Zhu, *Pharm. Res.* **2014**, *31*, 1002–1014.

[31] R. C. Braga, V. M. Alves, M. F. B. Silva, E. Muratov, D. Fourches, L. M. Lião, A. Tropsha, C. H. Andrade, *Mol. Inform.* **2015**, *34*, 698–701.

[32] J. A. Beltran, L. Aguilera-Mendoza, C. A. Brizuela, *BMC Genomics* **2018**, *19*, 672.

[33] A. J. Ruiz-Moreno, A. Reyes-Romero, A. Dömling, M. A. Velasco-Velázquez, *Molecules* **2021**, *26*, DOI 10.3390/molecules26071877.

[34] P. Polishchuk, *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.

[35] R. Rodríguez-Pérez, J. Bajorath, *Sci. Rep.* **2021**, *11*, 14245.

[36] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic Acids Res.* **2019**, *47*, D930–D940.

[37] T. Kalliokoski, C. Kramer, A. Vulpetti, P. Gedeck, *PLoS One* **2013**, *8*, e61007.

[38] G. A. Landrum, **2020**, DOI 10.5281/zenodo.3815117.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, . Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[40] S. M. Lundberg, S. I. Lee, in *Adv. Neural Inf. Process. Syst.* (Eds.: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., **2017**, pp. 4766–4775.

[41] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.

[42] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615–2623.

[43] M. D. Shultz, *J. Med. Chem.* **2019**, *62*, 1701–1714.

[44] P. G. Polishchuk, V. E. Kuźmin, A. G. Artemenko, E. N. Muratov, *Mol. Inform.* **2013**, *32*, 843–853.

[45] G. P. Wellawatte, A. Seshadri, A. D. White, *Chem. Sci.* **2022**, *13*, 3697–3705.

[46] R. Rodríguez-Pérez, J. Bajorath, *J. Med. Chem.* **2020**, *63*, 8761–8777.

[47] D. Jiang, Z. Wu, C. Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, T. Hou, *J. Cheminform.* **2021**, *13*, 1–23.

[48] R. F. Lameiro, A. Shamim, F. Rosini, R. Cendron, P. H. J. Batista, C. A. Montanari, *Future Med. Chem.* **2021**, *13*, 25–43.

[49] J. Lameira, V. Bonatto, L. Cianni, F. R. Rocho, A. Leitão, C. A. Montanari, *Phys. Chem. Chem. Phys.* **2019**, *21*, 24723–24730.

[50] A. I. Porrás, Z. E. Yadon, J. Altcheh, C. Britto, G. C. Chaves, L. Flevaud, O. A. Martins-Filho, I. Ribeiro, A. G. Schijman, M. A. Shikanai-Yasuda, S. Sosa-Estani, E. Stobbaerts, F. Zicker, *PLoS Negl. Trop. Dis.* **2015**, *9*, DOI 10.1371/journal.pntd.0003697.

[51] M. C. Field, D. Horn, A. H. Fairlamb, M. A. J. Ferguson, D. W. Gray, K. D. Read, M. De Rycker, L. S. Torrie, P. G. Wyatt, S. Wyllie, I. H. Gilbert, *Nat. Rev. Microbiol.* **2017**, *15*, 217–231.

[52] D. T. Stanton, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11–20.

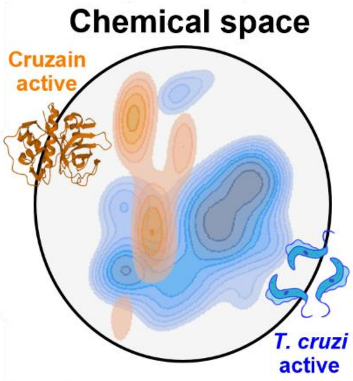
[53] P. Labute, *J. Mol. Graph. Model.* **2000**, *18*, 464–477.

[54] B. Zdrazil, R. Guha, *J. Med. Chem.* **2018**, *61*, 4688–4703.

[55] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[56] D. G. Silva, J. R. Gillespie, R. M. Ranade, Z. M. Herbst, U. T. T. Nguyen, F. S. Buckner, C. A. Montanari, M. H. Gelb, *ACS Med. Chem. Lett.* **2017**, *8*, 766–770.

[57] J. C. Gomes, L. Cianni, J. Ribeiro, F. R. Rocho, S. C. M. Silva, P. H. J. Batista, C. B. Moraes, C. H. Franco, L. H. G. Freitas-Junior, P. W. Kenny, A. Leitão, A. C. B. Burtoloso, D. de Vita, C. A. Montanari, *Bioorg. Med. Chem.* **2019**, *27*, 115083.



Graphical Abstract

The lack of translation of cruzain biochemical activity to the biological activity on *Trypanosoma cruzi* is investigated with machine learning. The figure shows a density plot on the t-SNE projection of the chemical space of active compounds (pIC_{50} > 7.0) exhibiting a low degree of overlap.