LEMYP
Laboratorio de Estructura Molecular y Propiedades

KHIPU LATIN AMERICAN MEETING IN ARTIFICIAL INTELLIGENCE
11-15 November, 2019 - Montevideo, Uruguay

CONICET
UNIVERSIDAD NACIONAL DEL NORDESTE
IQUIBA-NEA

PRELIMINARY RESULTS OF SUPERVISED MODELS TRAINED WITH CHARGE DENSITY DATA FROM CRUZAIN-INHIBITORS COMPLEXES

Villafañe, Roxana N.[a],Luchi, Adriano M.[a], Angelina, L. Emilio[a], Peruchena, Nélida M.[a]

[a] Lab. Estructura Molecular y Propiedades, IQUIBA-NEA, Universidad Nacional del Nordeste, CONICET, FACENA, Av.Libertad 5470, Corrientes 3400, Argentina.
E-mail: noelia0618@gmail.com

INTRODUCTION

Proteins are the most versatile biological molecules, with diverse functions.

Recently, the artificial intelligence (AI) community has turned attention to specific topics related to proteins as: protein folding [1], structural analysis [2], protein-ligand affinity estimation [3], among others. Cruzain (Cz) is a cysteine protease involved in Chagas disease with several Cz-inhibitor complexes deposited in the Protein Data Bank (PDB). Unfortunately, the number of structures solved-up to date is scarce for the requirements of a machine learning optimization algorithm. Another issue is the high dimensionality of the data involved in the structure-based approaches for drug design.

When high dimensions are presented, it is important to measure the feature importance and select the most discriminative features [4]. In this scenario, feature removal is important due to the overfitting problem associated with high dimensions. In this work, we optimized a support vector machine algorithm with recursive feature elimination (SVM-RFE) in order to compute the importance of features based on the weights of the model. SVM-RFE removes the weights with less importance, and the model is rebuilt [5].

Our goal is to take advantage of charge density data to find out favorable interactions (to stabilize the complex), which might explain the greater binding affinity of the more active inhibitors and the unfavorable (or less favorable) interactions that dominate the binding of the less active ones. Fig 1 shows that a comparative analysis of such intricate network of interactions for a set of Cz-Inh complexes cannot be performed by visual inspection of the molecular graphs by a human operator.

The present work is the first step for further analysis of topological data of Cz-ligand complexes under study. We hope that results will shed light to understand the inhibition mechanism of Cruzain.

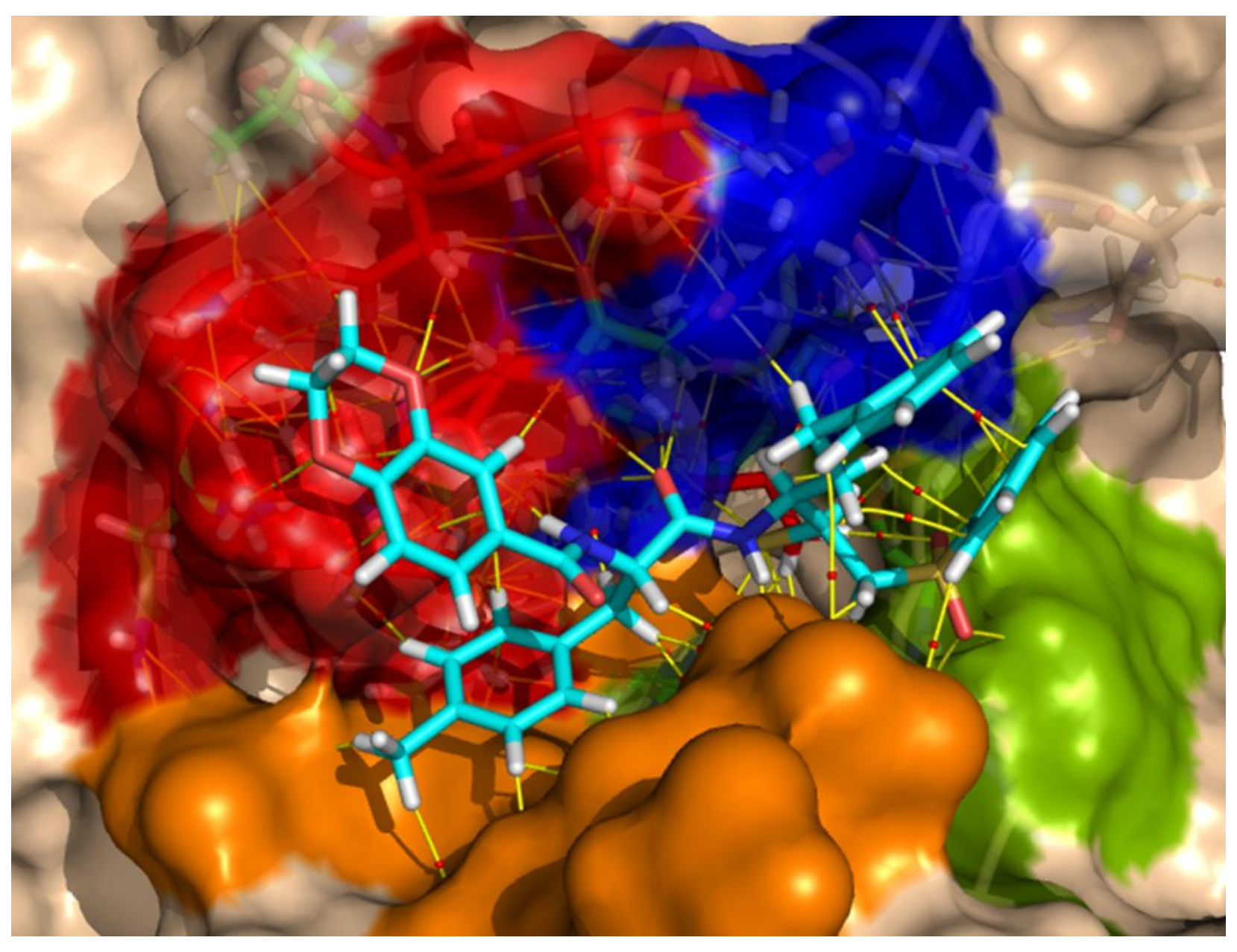

Figure 1. View of the intricate networks of interactions on the structure of the Cz-9d complex. Charge density topological elements describing the noncovalent interactions are depicted with small red circles (BCPs) and yellow lines connecting each BCP to both interacting atoms (BPs).

MATERIALS AND METHODS

Simulation Protocol

Jaishankar [6] synthesized and determined the inhibition constants against Cz of a series of vinyl sulfone analogues closely related to K-777. Although the experimental structure of these vinyl sulfone analogues in the complex with Cz has not been determined yet (except for K-777, pdb id = 2OZ2), for peptide-like Cz-inhibitors, a reasonably accurate initial guess of the inhibitor binding mode can be constructed "by hand" by placing each residue in the inhibitor sequence P1′, P1, P2, and P3 into its own enzyme subpockets S1′, S1, S2, and S3. Initial coordinates of the complex were taken from the structure of Cz bound to K-777 (pdb id = 2OZ2). By performing substitutions at P2 and P3 residues of K-777 to get the analogues reported by Jaishankar, closely related complexes were constructed and then refined by MD simulations. All the Cz−inhibitor complex simulations were carried out with Amber14 [7] software package at 300 K temperature and extended up to 50 ns overall simulation time in a truncated octahedral periodic box of TIP3P water molecules. Amber ff14SB forcefield was used for proteins residues. The antechamber software in the Amber-Tools package was used to generate ligand inhibitor parameters with GAFF forcefield and RESP charges.

Quantum Theory of Atoms in Molecules (QTAIM)

The structure of the potential energy minimum was selected from the MD trajectories of Cz−Inh complexes as a single representative structure upon which the charge density analysis was done. Because accurate quantum mechanical calculations are still forbidden for full biomolecular complexes, reduced models were constructed from the potential energy minimum structures. A total of 28 residues (~570 atoms) were included in the reduced models: the vinyl sulfone inhibitor and the surrounding residues in a spherical volume of about 5 Å centered on the inhibitor atoms. The charge density was computed by density functional theory methodology with the M06-2x dispersion corrected hybrid functional and 6-31G(d) as the basis set, as implemented in Gaussian 09 package [8]. The topological analysis of charge density was then performed with Multiwfn software [9].

Support vector machines – recursive feature elimination (SVM-RFE)

Charge density values associated to 319 noncovalent interactions per complex were used as features to train a linear SVM classifier. Therefore, in this article, we restricted ourselves to linear SVM because our main interest was in uncovering relationships between the features (i.e.,molecular interactions) and the biological activities to understand, ultimately, the enzyme inhibition mechanism.

SVM-RFE is a feature selection algorithm based on backward elimination of features with lowest weights. In each iteration, the SVM model is trained with the current subset of features, the weight (|**w**|) of each feature is calculated according to the SVM classifier, the features are ranked according to |**w**|, and then, the bottom-ranked features are eliminated. SVM-RFE was implemented by using the *scikit-learn* module of Python [10]

ACKNOWLEDGEMENTS

CONICET
SGCyT UNNE
NVIDIA

RESULTS AND DISCUSSION

In this work, we have applied support vector machine-recursive feature elimination (SVM-RFE) [11] to automate the process of extracting information from charge density molecular graphs and to exhaustively exploit the charge density data.

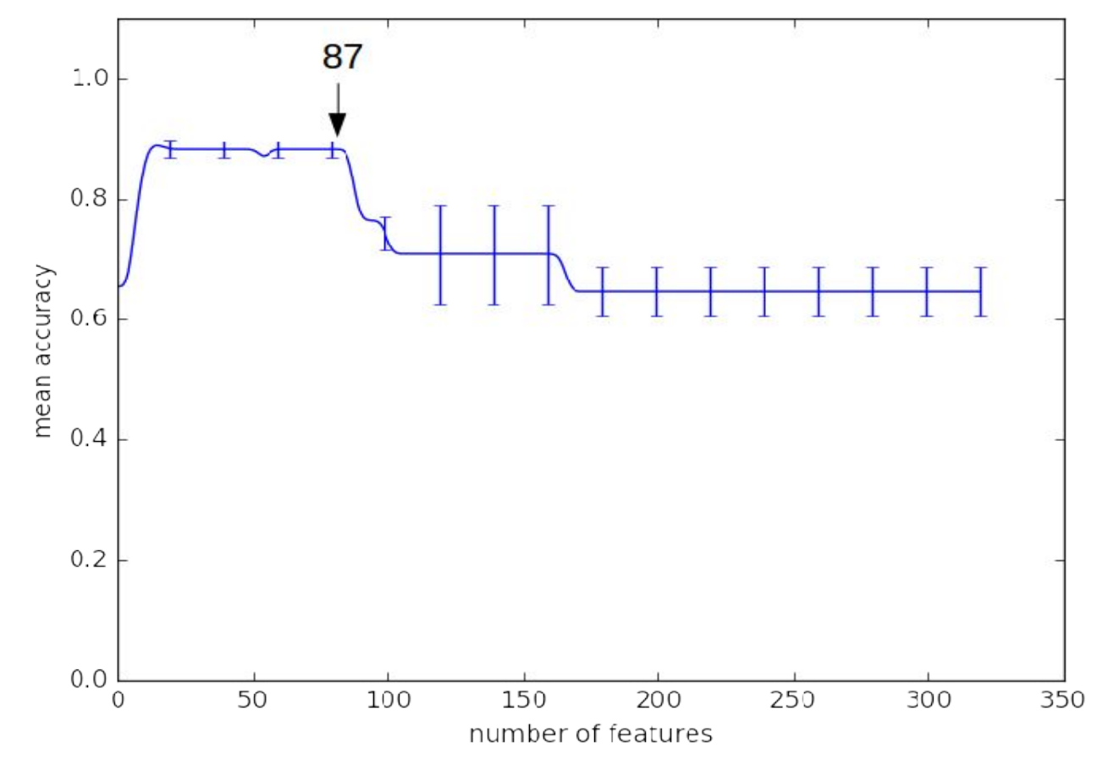

Figure 2. Iterative process of backward feature elimination and SVM model training with the remaining features. The mean accuracy of the SVM model is depicted as a function of the number of features. Error bars represent the variances of accuracy values among the folds.

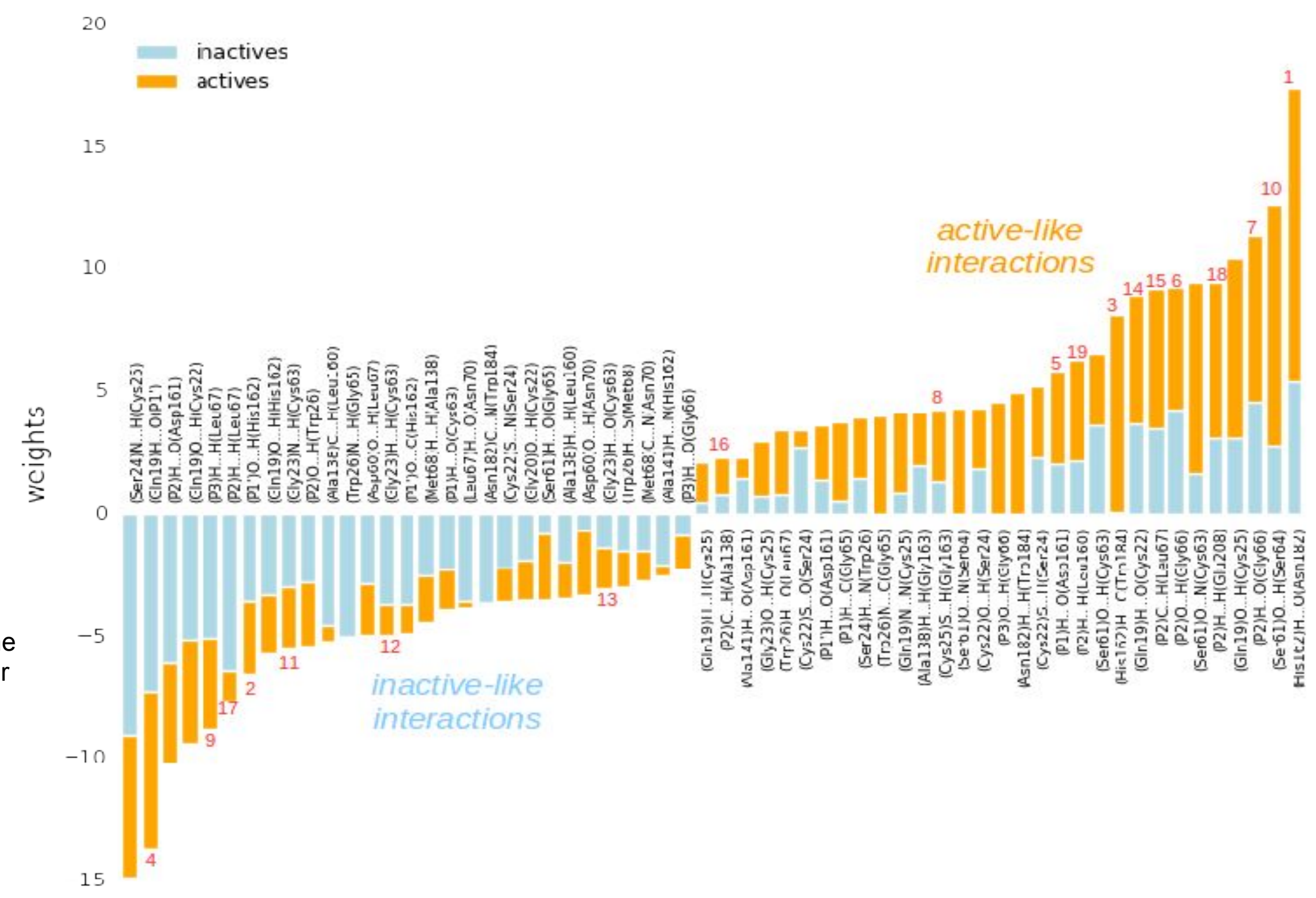

Figure 3. Top interactions (features) selected by the SVM model to make the class classifications. The numbers in red indicate the interactions
discussed in the text

SVM-RFE was built with a data set containing 319 interactions at the beginning, and then, the less relevant features were iteratively eliminated by a backward selection procedure [Fig 2]. Fig 3 shows the top interactions (features) that were used by the model to make the classification between active-like and inactive-like ligands. The total height of stacked bars represents the interaction importance for the classification task while each category within the bar represents the charge density contribution of the two classes (active and inactive in orange and light blue, respectively) to the overall feature importance. As can be seen in the figure, interactions with positive coefficients have overall greater contributions from compounds labeled as actives while the opposite is true for interactions with negative model coefficients, namely, their most important contributions come from compounds labeled as inactives.

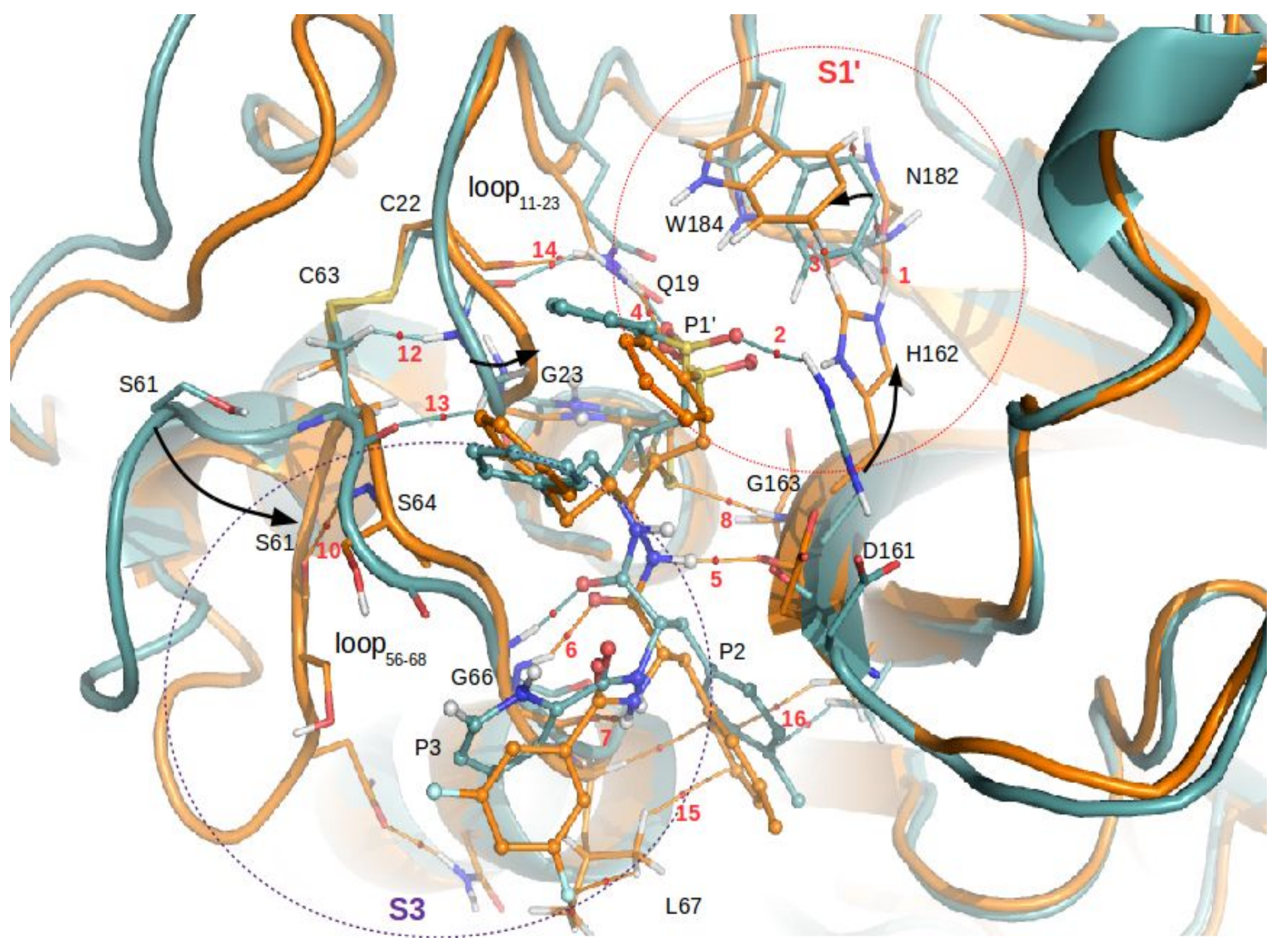

Figure 4. Structural superposition of Cz-6b (orange) and Cz-8d (light blue) complexes. Charge density topological elements for atomic interactions are also depicted: BPs connecting the nuclei are depicted in orange and light blue for Cz-6b and Cz-8d, respectively. BCPs are shown in small red spheres. Numbers in red indicate the most significant interactions (the same as Figure 3). Arrows indicate protein backbone displacement between Cz-8d and Cz-6b complexes.

Analysis of intermolecular interactions revealed that backbone−backbone hydrogen bonds between the peptide-like inhibitor and enzyme and interactions with the Leu67 residue play a key role in proper anchoring of the inhibitor to the Cz binding cleft. However, a quantitative structure−activity relationship could not be derived by considering only the intermolecular interactions between Cz residues and inhibitor atoms. On the other hand, if intramolecular contacts involving protein residues are also analyzed with the help of the SVM- RFE model, it becomes clear that a more indirect mechanism of enzyme inhibition involving extensive conformational changes within the protein structure operates under the hood.

CONCLUSIONS

By using a simple and interpretable linear SVM classification model coupled with an RFE procedure, it is possible to extract useful information about what are the most important interactions to discriminate between active and inactive (or less active) compounds against Cz. Our model allowed us to point out 19 inter-/intramolecular main interactions that could explain the principal changes in the complexes under analysis.

REFERENCES

[1] Evans R, Jumper J, Kirkpatrick J, Sifre L, Green TFG, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Jones DT, Silver D, Kavukcuoglu K, Hassabis D (2018) *De novo structure prediction with deep-learning based scoring.* A.W.Senior In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction

[2] Giulini M, Potestio R. (2019) *A deep learning approach to the structural analysis of proteins.* Interface Focus 9: 20190003.

[3] Zhang H, Liao L, Saravanan KM, Yin P, Wei Y. (2019) *DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity.* PeerJ 7:e7362

[4] Lin X, Li C, Zhang Y, Su B, Fan M, Wei H. (2018) *Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics.* Molecules 23(1) 52.

[5] Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models, 1st ed.; CRC Press: Boca Ratón, Florida. 2019.

[6] Jaishankar P, Hansell E, Zhao DM, Doyle P S, McKerrow JH, Renslo AR. (2008) *Potency and Selectivity of P2/P3-Modified Inhibitors of Cysteine Proteases from Trypanosomes.* Bioorganic & Medical Chemistry Letters 18, 624-628.

[7] Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. (2005) *The Amber Biomolecular Simulation Programs.* Journal of Computational Chemistry 26, 1668-1688

[8] Frisch et al *Gaussian 09*, Revision A.02; Gaussian, Inc., 2016

[9] Lu T, Chen F. (2012) *Multiwfn: A Multifunctional Wavefunction Analyzer.* Journal of Computational Chemistry 33, 580-592

[10] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. (2011) *Scikit-Learn: Machine Learning in Python.* Journal of Machine Learning. Research 12, 2825-2830.

[11] Guyon I, Weston J, Barnhill S, Vapnik V. (2002) *Gene Selection for Cancer Classification Using Support Vector Machines.* Machine Learning 2002, 46, 389-422.

Khipu 2019 Latin American Meeting in Artificial Intelligence