frontiers in Chemistry

ORIGINAL RESEARCH
published: 29 April 2020
doi: 10.3389/fchem.2020.00243

An Innovative Sequence-to-Structure-Based Approach to Drug Resistance Interpretation and Prediction: The Use of Molecular Interaction Fields to Detect HIV-1 Protease Binding-Site Dissimilarities

Nuno G. Alves[1†], *Ana I. Mata*[1†], *João P. Luís*[1†], *Rui M. M. Brito*[1,2] *and Carlos J. V. Simões*[1,2*]

[1] Department of Chemistry, Coimbra Chemistry Centre, University of Coimbra, Coimbra, Portugal, [2] BSIM Therapeutics, Instituto Pedro Nunes, Coimbra, Portugal

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Sinosh Skariyachan,
St. Pius X College, Rajapuram, India
Zhijun Li,
University of the Sciences,
United States

***Correspondence:**
Carlos J. V. Simões
carlos.simoes@bsimtx.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: 06 December 2019
Accepted: 13 March 2020
Published: 29 April 2020

Citation:
Alves NG, Mata AI, Luís JP, Brito RMM and Simões CJV (2020) An Innovative Sequence-to-Structure-Based Approach to Drug Resistance Interpretation and Prediction: The Use of Molecular Interaction Fields to Detect HIV-1 Protease Binding-Site Dissimilarities. Front. Chem. 8:243. doi: 10.3389/fchem.2020.00243

In silico methodologies have opened new avenues of research to understanding and predicting drug resistance, a pressing health issue that keeps rising at alarming pace. Sequence-based interpretation systems are routinely applied in clinical context in an attempt to predict mutation-based drug resistance and thus aid the choice of the most adequate antibiotic and antiviral therapy. An important limitation of approaches based on genotypic data exclusively is that mutations are not considered in the context of the three-dimensional (3D) structure of the target. Structure-based *in silico* methodologies are inherently more suitable to interpreting and predicting the impact of mutations on target-drug interactions, at the cost of higher computational and time demands when compared with sequence-based approaches. Herein, we present a fast, computationally inexpensive, *sequence-to-structure*-based approach to drug resistance prediction, which makes use of 3D protein structures encoded by input target sequences to draw binding-site comparisons with susceptible templates. Rather than performing atom-by-atom comparisons between input target and template structures, our workflow generates and compares Molecular Interaction Fields (MIFs) that map the areas of energetically favorable interactions between several chemical probe types and the target binding site. Quantitative, pairwise dissimilarity measurements between the target and the template binding sites are thus produced. The method is particularly suited to understanding changes to the 3D structure and the physicochemical environment introduced by mutations into the target binding site. Furthermore, the workflow relies exclusively on freeware, making it accessible to anyone. Using four datasets of known HIV-1 protease sequences as a case-study, we show that our approach is capable of correctly classifying resistant and susceptible sequences given as input. Guided by ROC curve analyses, we fined-tuned a dissimilarity threshold of classification that results in remarkable discriminatory performance (accuracy ≈ ROC AUC ≈ 0.99), illustrating the high potential of *sequence-to-structure*-, MIF-based approaches in the context of drug

resistance prediction. We discuss the complementarity of the proposed methodology to existing prediction algorithms based on genotypic data. The present work represents a new step toward a more comprehensive and structurally-informed interpretation of the impact of genetic variability on the response to HIV-1 therapies.

Keywords: drug resistance prediction, Molecular Interaction Fields, *sequence-to-structure* algorithm, binding-site dissimilarities, HIV-1 protease

INTRODUCTION

Drug resistance is one of the greatest threats of the twenty first century. Fundamentally, the problem resides in the development and spread of resistance-conferring mechanisms among infectious pathogens such as viruses and other microbial targets (McKeegan et al., 2002). Importantly, the selection of random mutations stands out as one of the main mechanisms of acquiring resistance, particularly relevant in viruses which mutate at high frequencies. RNA viruses, for instance, have a mutation rate estimated at 10^{-4} per nucleotide per replication, while DNA viruses have a rate of 10^{-8} per nucleotide per replication (Vere Hodge and Field, 2011; Mason et al., 2018). The extreme variability and rapid mutational spectrum of viral genomes, ongoing viral replication, and prolonged drug exposure linked with the selection and widespread of new drug-resistant strains is still a matter of great concern and importance, particularly in immunocompromised populations (Strasfeld and Chou, 2010; Mason et al., 2018). While a limited number of antiviral drug classes are getting approved for human use, an increasing resistance to some of the most effective available antivirals for HIV/AIDS, herpes, influenza and hepatitis, is being observed. Furthermore, the unpredictability of viral evolution and drug resistance means that antiviral treatments remain costly to the health care systems and are still associated with a significant risk of mortality, particularly in low- and middle-income countries (Irwin et al., 2016). Hence, *a priori* understanding and prediction of resistance against drug targets is of paramount importance toward developing more effective and longer lasting treatment options and regimens.

Antiviral drug resistance has been extensively studied in the rapidly mutating human immunodeficiency virus (HIV). HIV-1, in particular, is one of the most studied virus and the increasingly affordable and accessible genotypic data from clinical HIV-1 strains, together with corresponding data on strain susceptibility or resistance toward several drugs, have sparked the development of several genotypic interpretation systems for prediction of phenotypic drug resistance and therapy response based on genotype (Bonet, 2015). Said systems include (a) rule-based algorithms, including the *Agence Nationale de Recherche sur le Sida* (ANRS) (Brun-Vézinet et al., 2003), the Stanford HIV Drug Resistance Database interface (HIVdb) (Tang et al., 2012), Rega (Van Laethem et al., 2002), and HIV-GRADE (Obermeier et al., 2012a), which heavily rely on the periodic update of mutation-resistance profile lists, and on the knowledge of expert panels; and (b) machine learning-based algorithms trained on large sets of genotype–phenotype pairs to predict the *in vitro* resistance to a specific drug, with renowned examples such as *geno2pheno* (Beerenwinkel et al., 2003) and SHIVA (Riemenschneider et al., 2016). These sequence-based methods are relatively fast and low cost, justifying their routine use to support medical decision in HIV pharmacotherapy (Vercauteren and Vandamme, 2006).

The most relevant computational predictors of antiviral drug resistance currently available share the shortcoming of being purely based on genotypic sequence data. By disregarding the three-dimensional structural context and enzymatic function of the mutated amino acid residues, these systems fail to capture the links between genetic viral mutations and the corresponding mutation-induced structural changes to the effector protein viral machinery (Cao et al., 2005; Weber and Harrison, 2016; Khalid and Sezerman, 2018). This means that such methods are limited in their predictive power and interpretability toward novel mutations and combinations of mutations that go beyond the information accessible for training, such as mutation patterns that are encountered in only a small number of patients.

In contrast, structure-based methods hold potential to help understanding and eventually predicting resistance mechanisms for previously unknown data, shedding light on the elusive link between novel mutations and drug resistance. This may be justified by the fact that such methods can take advantage of available structural information on protein-ligand complexes and structural modeling of point mutations in the protein structure (Hao et al., 2012). Reported examples of the use of structure-based methods include the application of molecular docking to predict resistance or susceptibility of HIV1-PR to different inhibitors (Jenwitheesuk and Samudrala, 2005; Toor et al., 2011), the use of molecular dynamics simulations to study the impact of mutations on enzyme dynamics, stability and binding affinity (Hou and Yu, 2007; Agniswamy et al., 2016; Sheik Amamuddy et al., 2018), and the use of computational mutation scanning protocols to extract insights on free energy and binding affinity changes resulting from active site and non-active site mutations (Hao et al., 2010). Even though these methods are constantly adding new pieces to the puzzle and opening opportunities in the understanding of drug resistance, they suffer from various drawbacks, such as being time-consuming and offering limited predictive accuracy. As a result of such limitations, the primary challenge facing structure-based drug resistance prediction is to achieve an acceptable balance between prediction accuracy and computational efficiency to become both reliable and fast tools to be used in clinic context (Hao et al., 2012). In fact, some of the most recent reports describe the use of machine learning strategies merging both sequence and structural data in attempt

to achieve such balance (Masso and Vaisman, 2013; Yu et al., 2014; Khalid and Sezerman, 2018).

In this contribution, we describe a fast, computationally inexpensive, *sequence-to-structure*-based approach to the prediction of drug resistance. The proposed workflow makes use of an archetypal GRID-based method (Goodford, 1985) involving the generation and comparison of Molecular Interaction Fields (MIFs). MIFs may be defined as the spatial variation of interaction energies between a molecular target structure and selected types of chemical probes laid out on a three-dimensional (3D) grid (Cruciani, 2005). The broad range of applications of MIFs extends from ligand-based methodologies, e.g., 3D Quantitative Structure-Activity Relationships (3D-QSAR) models, drug metabolism and pharmacokinetics (DMPK) predictions and pharmacophore elucidation, all the way to structure-based drug design, including binding site detection and molecular docking (Artese et al., 2013). Within the context of viral drug resistance, MIFs hold potential in capturing subtle, mutation-induced, chemical perturbations within the binding site of resistant or susceptible viral structures, thus representing a promising approach to anticipating the impact of mutations on the response to antiviral drugs with atomistic detail.

HIV-1 protease (HIV1-PR) is one of the most characterized viral enzymes, with extensive structural, inhibitor, and mutation data available (Weber and Agniswamy, 2009). As of late 2019, the RCSB Protein Data Bank (RCSB PDB, 2000) ranks HIV-1 as the virus holding the highest number of available structures (2,586), majorly obtained through X-ray crystallography. Of these, the PDB returns 662 entities with at least 90% identity to the HIV1-PR subtype B *consensus* sequence from a BLAST sequence search (Stanford University, 1998a). The search by *consensus* sequences of other HIV-1 subtype B enzymes (Stanford University, 1998a) returns 586 structures for reverse transcriptase and 190 for integrase. With such amount of structural information available, we have built the framework of the present work using HIV1-PR as our first case-study. Commercially available HIV-1 protease inhibitors (PIs) are competitive peptidomimetics with a core structural scaffold that mimics the tetrahedral transition state of HIV1-PR substrate. Although these drugs are chemically distinct, their active conformations are superimposable, and generally establish the same pharmacophoric interactions with their target (Wlodawer and Erickson, 1993; King et al., 2004; Qiu and Liu, 2011; Nayak et al., 2019). Many mutations in HIV1-PR translate into changes in the structure and binding site physicochemical environment, thus affecting the affinity of PIs and representing a hurdle to achieving long-term viral suppression (Irwin et al., 2016; Pawar et al., 2019; Wensing et al., 2019). A quantitative analysis of HIV1-PR drug-resistant mutation frequency, with particular focus on the binding site, was performed using public sequence datasets to support the potential of a MIF-based approach to capturing mutation-induced active site dissimilarities. From this perspective, the workflow proposed here encompasses the use of a conservative structural modeling step for the generation of a HIV1-PR structure from its respective amino acid sequence, and a MIF-based structural alignment and chemical dissimilarity detection step comparing the input *sequence-structure* pair with a carefully selected naïve, susceptible template *sequence-structure* pair. We demonstrate that the quantification of such dissimilarity, depicting the extent of structural, physicochemical and pharmacophoric alterations introduced by mutations, allows for an accurate prediction of HIV1-PR's resistance to PIs.

Compared with previous approaches reported in the literature, and to the best of our knowledge, this work stands out as a first implementation of a fast, *sequence-to-structure*-based algorithm capable of discriminating susceptible and resistant HIV1-PR sequences. Considering that the problem of mutation-induced resistance cuts across virtually all infectious diseases, we believe the approach reported herein may be extended to a wide range of microbial targets besides HIV-1, thus helping rationalize and personalize the therapeutic decision-making process.

MATERIALS AND METHODS

The availability of a public and curated database such as HIVDB (Stanford University, 1998c; Rhee et al., 2003) allows access to HIV1-PR sequences with known levels of resistance, and thus to establish datasets for the development of new methodologies to predict HIV1-PR resistance to protease inhibitors (PIs). This section describes the materials and methods employed in (1) the preparation of sequence datasets with various levels of resistance to PIs; (2) frequency analysis of *major* and *minor* mutations in the sequence datasets in (1); (3) the structural modeling of the reference structure used as template for subsequent modeling of HIV1-PR structures corresponding to each sequence in the datasets; (4) the core components of the proposed algorithm, including the calculation and comparison of pairwise Molecular Interaction Field points between the resulting structural models and the selected naïve template structure; and (5) the performance metrics used to test and evaluate the predictive power of the developed structure-based drug-resistance classification algorithm. A general workflow illustrating (4) and (5) is sketched (draw.io, 2005) in **Figure 1** and the complete script *HIV1predict.sh* for running the sequences is available at GitHub (Alves et al., 2019b). Calculations were run on a 64-bit CentOS 6 Linux server with an Intel Xeon CPU (E5620) at 2.40 GHz (further information as **Supplementary Table S1**).

Datasets of Resistant and Susceptible Sequences

A set of genotype-phenotype correlated HIV1-PR sequences was retrieved from HIVDB, version 8.7 (Stanford University, 1998b,c), and filtered by drug class for PIs. The considered PIs include darunavir, fosamprenavir, atazanavir, indinavir, lopinavir, nelfinavir, saquinavir, and tipranavir. Analyzing the subtype B HIV1-PR sequence of each isolate, i.e., a viral sample obtained from an infected individual, and considering positions with a mixture of amino acids, all possible mutation patterns were written to the FASTA format using a script written in-house (Alves et al., 2018f).

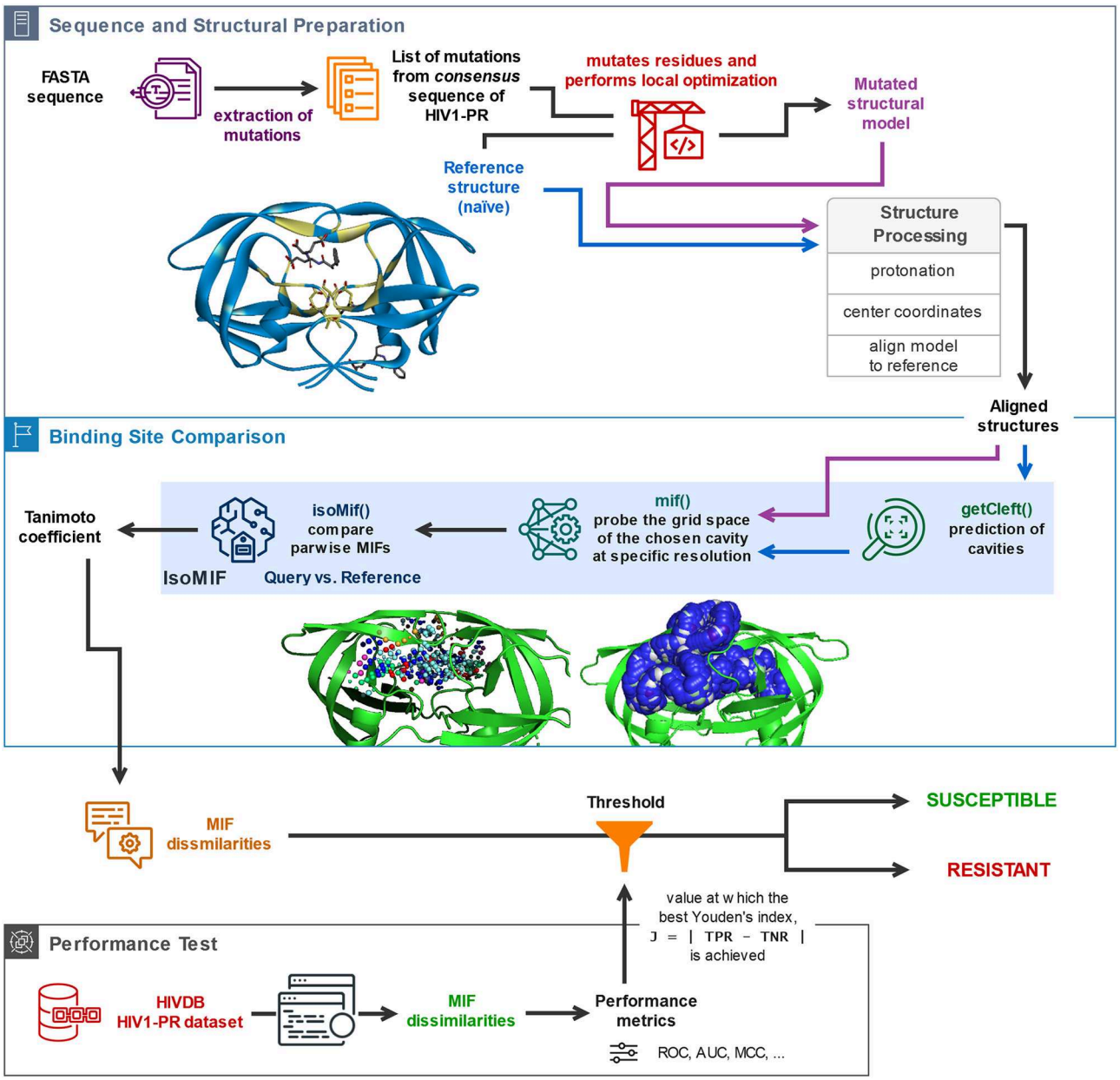

FIGURE 1 | General description of the workflow underlying the proposed *sequence-to-structure*-based drug resistance classifier, holding platform applications to multiple microbial targets. The depicted algorithm starts off with the sequential reading of an amino acid sequence in FASTA format, followed by identification of the present mutations by comparison with the consensus sequence, insertion of the identified mutations in the naïve template, and processing of the structural models for alignment and comparison. Upon structural alignment of the target or database (predicted) structure with the naïve template structure, six types of MIF probe points are computed for the former structure and superimposed with pre-calculated MIF probe points of the latter structure. Calculation of MIF dissimilarities by means of a Tanimoto coefficient proceeds. The bottom panel represents the process of performance evaluation of the proposed classifier based on its application to a large dataset of *sequence-structure* pairs generated for HIV1-PR sequences retrieved from HIVDB. Included are performance metrics such as accuracy, Matthews Correlation Coefficient (MCC), and the area under the receiver-operating-characteristic curve (ROC AUC).

The genotype-phenotype correlation results from the *in vitro* PhenoSense assay (Zhang et al., 2005), which measures the levels of resistance to a PI compared to the wild-type sequence. Following the categorization of susceptibility to PIs described by Rhee et al. (2006), the collected sequences were classified as follows:

- *Susceptible*. Sequences holding <3.0-fold resistance to all PIs in the dataset were considered susceptible (N = 7,768) [Susceptible].
- *Resistant*. Sequences holding more than 20.0-, or 15.0-, or 10.0-fold resistance to all PIs, resulting in three resistant subgroups of increasing degree of resistance: respectively, [Res_{20}] (N =

60) [Res20], [Res_{15}] ($N = 83$) [Res15], which encompasses [Res_{20}] plus 23 sequences holding between 15- and 20-fold resistance, and [Res_{15}] ($N = 873$) [Res10], which encompasses [Res_{20}] and [Res_{15}] plus 790 additional sequences holding between 15- and 10-fold resistance.

Counting of Mutations in HIV1-PR

The quantification of *major* and *minor* mutations (Weber and Agniswamy, 2009) in all datasets was carried out using scripts written in-house (Alves et al., 2018a,b, 2019a) that sequentially read the listing of mutations for each sequence, extract either the *major* or *minor* mutations, and count them for each sequence. Said script was applied to quantify *major* and *minor* mutations in the HIV1-PR binding site.

Preparation of HIV1-PR Structures

Using PDB's BLAST utility (Altschul et al., 1990) to guide the choice of a template for homology modeling, a sequence search, with a 10.0 E-value cut-off and at least 50% identity to the HIV1-PR subtype B *consensus* (Stanford University, 1998a), resulted in 784 entities available. With a more refined query of at least 95% identity to the HIV1-PR subtype B *consensus*, there were still 376 structures available to work with.

Out of these 376 structures, PDB entry 1NH0 for HIV1-PR was chosen as template structure for homology modeling by using PDB's BLAST utility (Altschul et al., 1990). It returned an E-value of 7.20281E-51, but since the intended work was heavily based on structure, our choice was also based on having the best resolution possible. The structure of 1NH0 holds 99% sequence identity (98/99) with the *consensus* B amino acid sequence of protease, HXB2 (henceforth referred to as *consensus sequence*), with one single mutation at position 37 (S37N), has 100% coverage of the sequence, and has been determined at 1.03 Å X-ray resolution. Importantly, this HIV1-PR sequence is known to be susceptible to all PIs.

In this work, Modeler version 9.19 (Šali and Blundell, 1993; Šali, 2019a) was used for predictive modeling of all HIV1-PR structures from their respective sequences. The listing of mutations present in each sequence was automated by scripting (Alves et al., 2018c) and followed by sequentially running the *mutate_model.py* script provided with Modeler (Šali, 2019b) to obtain the correct pattern of mutations and outputting the respective structural model. The procedure implemented in *mutate_model.py* performs local optimization of the mutated residues region and ensures that the obtained structural models are comparable to the template structure. The PDB structure itself (1NH0) was subjected to *mutate_model.py* in order to reverse the mutation present in the template with 99% identity (Asn37, on the outside of the protease) and keep on the *consensus sequence*, remove *HETATM* entries and *alt-locs*—thus yielding the reference template structure. This reference structure was used as template for the generation of the respective structural model of each input FASTA sequence present in the datasets.

All generated structural models were protonated using *Reduce*, version 3.23 (Word et al., 1999). The reference structure was centered to the origin of the axes of the cartesian coordinate system using *VMD*, version 1.9.3 (Humphrey et al., 1996). Structural alignment of all query models onto the centered reference structure was performed with *LovoAlign*, version 16.342 (Martínez et al., 2007).

Workflow for Detection and Scoring of Molecular Interaction Field Dissimilarities

The MIF module of the software package IsoMIF, version dated March 2015 (Chartier and Najmanovich, 2015), was used to generate Molecular Interaction Fields (MIFs) within the HIV1-PR binding sites. MIF-based alignment and calculation of pairwise MIF dissimilarities between reference and dataset binding sites proceeded using the IsoMIF module of the same package. The IsoMIF setup comprises three sequential modules: GetCleft, MIF, and IsoMIF.

Cavity Detection (GetCleft Module)

GetCleft (Gaudreault et al., 2015) was employed to predict cavities in the structure of the reference HIV1-PR (Alves et al., 2018e). This geometry-based method detects cavities by insertion of spheres of radius *r* between the non-hydrogen protein atoms, reducing such radius if they intersect with any neighboring atoms (clefts defined by the union of overlapping spheres). First, the top five largest cavities were searched at the same time, with a minimum and maximum sphere radius of 1.5 and 4.0 Å, respectively. The largest predicted cavity was visually confirmed to be completely enclosed within the HIV1-PR binding site, using *VMD*, version 1.9.3 (Humphrey et al., 1996). Next, such cavity volume represented by spheres was used to define the location of MIF interaction vectors to be calculated for the reference and all 3D HIV1-PR structural models.

Generation of Molecular Interaction Field (MIF) Probe Points (MIF Module)

The MIF module of IsoMIF was used to compute molecular interaction fields (MIFs) for six different chemical probe types (**Figure 2**): hydrophobic, aromatic, H-bond donor, H-bond acceptor, positive charge and negative charge. The pharmacophoric features shared by PIs (Wlodawer and Erickson, 1993; Nayak et al., 2019) highlight the importance of a conserved physicochemical environment in the binding site. Alterations of this environment are detected with the MIF probes (circled in **Figures 2A,B**) which allow for a quantification of changes caused by the presence of mutations. In this work, a grid resolution of 1.5 Å was defined to calculate the MIFs on the cleft covering the volume of the binding site. Such resolution was selected upon testing to achieve an adequate balance between speed and accuracy of IsoMIF pairwise field dissimilarity calculations.

Alignment of MIF Probe Points and Calculation of Dissimilarities (IsoMIF Module)

Field similarities were computed using the IsoMIF module, which employs a clique-based graph matching approach based on the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973) to perform functional alignments between the probe points under comparison. A grid spacing of 1.5 Å, a geometric distance threshold of 1.0 Å and a maximum of 100 cliques were used as parameters for the calculation of similarities between the

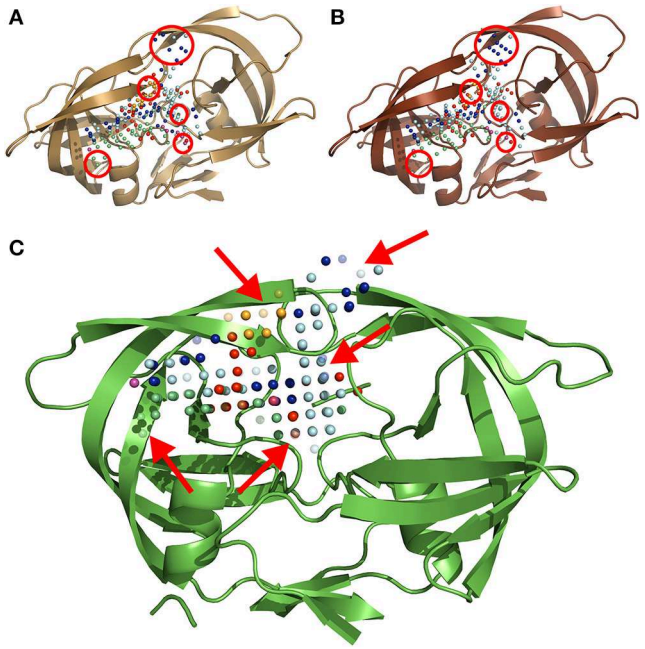

FIGURE 2 | A three-dimensional ribbon depiction of the quaternary structure HIV1-PR dimeric unit, including the six different types of MIF probe points calculated on the enzyme's binding site. (A) Structure and probe points of the reference naïve template. (B) Illustration of the structure and probe points of a mutated model resulting from a sequence holding over 20-fold resistance for PIs. (C) An exemplary output of IsoMIF, highlighting a MIF-based alignment and comparison of both structures: (A,B). The red circles in (A,B) denote areas where the difference in probe points is most visible. The arrows in (C) point to semi-transparent probes, representing the probes which are not shared between the two structures. Legend for the six (6) probe types legend: hydrophobic in light blue, aromatic in orange, H-bond donor in dark blue and acceptor in red, positive in green and negative charge in purple.

binding site of reference and structural models of HIV1-PR. Such similarities were then quantified by the Tanimoto coefficient (T_c), calculated as in Equation 1:

$$T_c = \frac{N_C}{N_R + N_Q - N_C} \tag{1}$$

where N_c is the number of common probe points to the two MIF maps under comparison; N_r and N_q represent the number of probe points present in the reference and query structure, respectively (**Figure 2C**) (Chartier and Najmanovich, 2015). The measurement of dissimilarity (Equation 2) between binding sites is justified by the fact that the focus of this work is set on the discrimination of resistant structures, when compared with a susceptible reference. Therefore, the chosen metric was dissimilarity rather than similarity:

$$dissimilarity\ coefficient = 1.0 - T_c \tag{2}$$

Analysis of Mutation Patterns Across Thousands of HIV1-PR Sequences

Analyses of the number and position of mutations were performed on HIV1-PR sequences in order to obtain information supporting and justifying the development of a *sequence-to-structure*-, MIF-based approach to antiviral resistance classification and prediction.

R version 3.4.3 (R Core Team, 2018) was used to conduct the analysis and generating the associated graphical representations. The R packages used in this work were ggplot2 (Wickham, 2009), *gplots* (Warnes et al., 2019), and ROCit (Khan and Brandenburger, 2019).

"Outlier" Detection on Binding-Site MIF Dissimilarities

Tukey's method (Tukey, 1949; Hoaglin, 2003), also referred to as Tukey's fences method, was used to detect outliers in the binding-site MIF dissimilarities results. Tukey's method is a statistical approach used to determine whether a value should be considered an outlier or not: the method relies on the interquartile range (IQR) measurement, which is calculated by the difference between the first quartile (Q1) and the third quartile (Q3) (see Equation 3). Q1 stands for the value in the dataset that holds 25% of the values below it and Q3 is the value in the dataset that holds 25% of the values above it.

$$IQR = Q3 - Q1 \tag{3}$$

According to Tukey's method, a value is considered an outlier if it is observed in the range described in Equation 4:

$$\begin{aligned} outlier < Q1 - k \times IQR \vee outlier > Q3 + k \times IQR \\ outlier < LowerBound \vee outlier > UpperBound \end{aligned} \tag{4}$$

where $k = 1.5$ indicates an outlier and $k = 3$ indicates an extreme outlier. For the purpose of the present work, only extreme outliers were discarded.

Evaluation of the Algorithm's Predictive Performance

The performance of our method at discriminating resistant from susceptible models was assessed by calculation of several metrics typically employed in the fields of predictive modeling and machine learning, particularly in cases where binary classification occurs. These included the Receiver Operating Characteristic (ROC) and the respective Area Under the Curve (ROC AUC). The ROC curve is a graphical representation of the True Positive Rate (TPR) as a function of the True Negative Rate (TNR), i.e., at various cut-off settings. The TPR is also known as Sensitivity (Equation 5), which measures the proportion of positive cases. On the other hand, the TNR is also calculated as 1—Specificity (Equation 6) and measures the proportion of true negative cases.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where *TP* represents the number of correctly identified resistant structures (true positives), *TN*, the number of correctly identified susceptible structures (true negatives), *FP*, the number of susceptible incorrectly predicted as resistant (false positives), and

FN the number of resistant incorrectly predicted as susceptible (false negatives).

Additional performance metrics included Accuracy (Equation 7) and Matthews Correlation Coefficient (MCC; see Equation 8) (Matthews, 1975; Florkowski, 2008; Powers, 2011).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \quad (8)$$

The dissimilarity threshold used for classification in resistant or susceptible *sequence-structure* pairs was derived from ROC curves, corresponding to the highest Youden's index (Youden, 1950), J, calculated as in Equation 9:

$$J = Sensitivity + Specificity - 1 \quad (9)$$

This index defines the maximum potential effectiveness of a classifier. It can be determined for all points of an ROC curve, although its maximum value represents the classifier optimal differentiating ability cut-point when equal weight is given to Sensitivity and Specificity (Ruopp et al., 2008).

RESULTS AND DISCUSSION

In this work, we describe a *sequence-to-structure*-, MIF-based method to assess binding-site dissimilarities across *sequence-structure* pairs, with the aim of predicting antiviral resistance—and using HIV1-PR as a case-study. It is generally accepted that the majority of resistance-conferring mutations occur in the binding site regions of viral enzymes (Weber and Agniswamy, 2009; Weber and Harrison, 2016). In order to further support the *rationale* and underlying assumptions of the proposed approach, we performed analysis of *major* and *minor* mutations of HIV1-PR binding site residues focusing on sequences known to be fully resistant and fully susceptible. For the sake of comparison, the quantification of mutations was also extended to *major* and *minor* mutations occurring in the remainder residues, i.e., residues not comprising the binding site region of HIV1-PR.

Counting of PI-Resistant Mutations in HIV1-PR Sequences

Resistance to PIs develops upon accumulation of mutations that increasingly impact the structure of HIV1-PR, resulting in highly-resistant variants of HIV-1. As mentioned by Weber and Agniswamy (2009), PI resistance is linked to the occurrence of primary (*major*) mutations, commonly associated with the active site where HIV PIs typically bind, resulting from structural changes that disrupt the van der Waals contacts and/or hydrogen bonding patterns in the inhibitor-protein interaction and promote direct steric hindrance, by altering the pocket volume or its physicochemical environment. Secondary (*minor*) mutations occur in addition to *major* mutations, acting like accessory mutations that compensate the flaws produced by *major* mutations and enhancing the resistance level (synergistic effect). Being less obvious, they seem to affect HIV1-PR catalysis, dimer stability, inhibitor binding kinetics, and/or active site re-shaping through long-range structural perturbations (Weber and Agniswamy, 2009; Weber and Harrison, 2016).

Our workflow follows a *sequence-to-structure* approach in attempt to capture changes to the structural and physicochemical determinants of HIV1-PR's binding site upon mutation, based on the assumption that these changes represent the main driver of antiviral resistance. To support this assumption, quantification of mutations known to contribute to PI resistance was carried out across the retrieved datasets. The version 8.7 HIVDB (Stanford University, 1998b,c,d) listed the following PI-resistant mutations for HIV1-PR:

- *Major* mutations: D30N, V32I, L33F, M46IL, I47VA, G48VM, I50VL, I54VTALM, L76V, V82AFTSL, I84V, N88SD, and L90M;
- *Minor* mutations: L10FIVRY, V11IL, K20RIMTV, L23I, L24IFM, M36I, K43T, M46V, G48ASTQL, F53LY, I54ST, Q58E, A71VTIL, G73STCADV, T74PS, V82MC, N83DS, I84AC, I85V, N88TG, and L89VT.

Even though not all sequences exhibit the same degree of resistance to each PI, we selected these two groups of *major* and *minor* PI-resistant mutations and quantitatively characterized their presence in our subsets. Since all HIV1-PR sequences in our dataset were retrieved from the same unique source, HIVDB (Stanford University, 1998c; Rhee et al., 2003), the percentage of sequences holding PI-resistant mutations distributed across the entire HIV1-PR sequence, as well as the percentage of PI-resistant mutations manifesting in residues comprising the binding site of HIV1-PR, were determined and compared among all four subsets: [Susceptible], [Res_{10}], [Res_{20}], and [Resistant*]–as represented in **Figure 3**.

Figures 3A,B shows that, as expected, all HIV1-PR sequences belonging to the *Susceptible* subset hold much less PIs-resistant mutations than those belonging to the *Resistant* subsets. The majority (98.24%) of susceptible HIV1-PR sequences does not hold any *major* mutations, while 1.74% contain one *major* mutation, and only one sequence (0.01%) comprises three *major* mutations. The presence of *major* mutations across drug-resistant sequences is higher, ranging from three to seven *major* mutations, implying that among these subsets the *major* mutations appear in the shape of mutation patterns rather than individual mutations. The presence of *minor* mutations (**Figure 3B**) follows a similar trend to that witnessed for *major* mutations, with susceptible sequences denoting a lower number when compared to their resistant counterparts. Approximately 98.25% of the susceptible sequences present two or less *minor* mutations, with about half of susceptible HIV1-PR sequences (53.3 %) displaying no *minor* mutations.

When comparing susceptible vs. drug-resistant sequences, it can be observed that resistance against PIs is linked to the presence of *major* mutations, as implied above (Weber and Harrison, 2016). However, within the subsets of drug-resistant sequences, a direct relation between the number of *major* mutations and the increase of resistance is not observed. Drug-resistant sequences show a higher frequency of *minor*

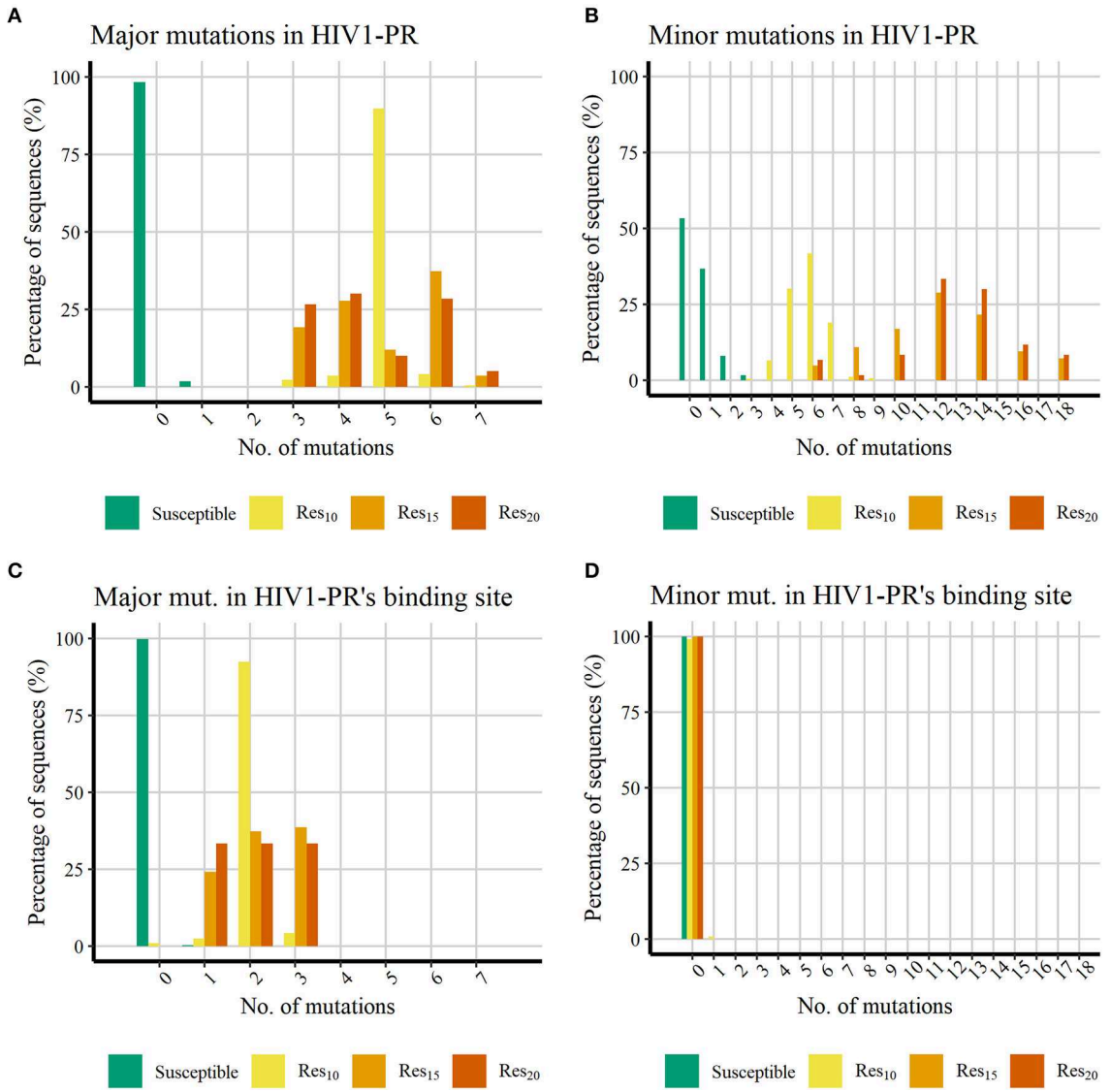

FIGURE 3 | Histograms representing the percentage of PI-resistant mutations in the datasets showing increasing number of PI-resistant mutations in the datasets retrieved from HIVDB. (A) Percentage of major mutations in the whole HIV1-PR. (B) Percentage of minor mutations in the whole HIV1-PR. (C) Percentage of major mutations in the HIV1-PR binding site. (D) Percentage of minor mutations in the HIV1-PR binding site. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

mutations, ranging from three to 18, with a visual apparent difference between sequences with lower resistance ([Res_{10}]) and the more resistant sequences ([Res_{15}] and [Res_{20}]). In [Res_{10}], 98.2% of the sequences have up to seven *minor* mutations, while 78.3% in [Res_{15}] and 93.3% [Res_{20}] have more than eight *minor* mutations. This trend in the profile of mutation distribution among the resistant sequences is in line with *minor* mutations acting as accessory mutations, appearing as patterns and not as individual mutations, and showing a similar trait as the one observed for the distribution of *major* mutations.

Analysis of *major* mutations located in HIV1-PR's binding site residues (**Figure 3C**), corresponding to sequence positions 30, 32, 47, 48, 50, 82, and 84, shows that 99.78% of the susceptible sequences do not display *major* mutations, while the remainder show only one *major* mutation. In contrast, less than 1% of resistant sequences lack *major* mutations in the drug binding site. Interestingly, the eight sequences representing this small fraction (0.91%) belong to the lower (10-fold) resistance subset ([Res_{10}]). All remaining drug-resistant sequences hold from one to three *major* mutations in the enzyme's binding site.

Counting of mutations in binding site residues of HIV1-PR exposes a systematic presence of *major* mutations in resistant HIV1-PR sequences, while also highlighting the absence of such mutations on 99.78% of their susceptible counterparts. This contrasting trait observed between the binding site region of susceptible and resistant HIV1-PR supports the development

of a structure-based drug-resistance classifier focusing on the detection and quantification of binding site dissimilarities.

Regarding the distribution of *minor* mutations across binding site residues, as represented in **Figure 3D**, mutations localized in sequence positions 23, 48, 82, and 84 were quantified among both HIV1-PR susceptible and drug-resistant sequences, revealing that the great majority does not present *minor* mutations in their respective binding sites. Only a small percentage of susceptible (0.01%) and resistant sequences (0.91%) show *minor* mutations in this region. It should be noted that the small subset of resistant sequences holding a *minor* mutation in their binding site region correspond to sequences that do not display *major* mutations in the active site.

These results show that the binding site *minor* mutations are uncommon on the datasets of HIV1-PR sequences—be they resistant or susceptible. Although such mutations appear to be important to increase the enzyme resistance's by stabilizing the mutated protein structure, they seem to produce limited direct effect on the enzyme's binding site, where they are mostly absent. Thus, these results seem to be in agreement with our motivation to explore a quantitative detection of binding-site dissimilarities to predict HIV1-PR resistance to PIs, as the *major* mutations play the main role on altering the binding site conformation, volume and/or physicochemical environment.

The quantification of mutations in the datasets retrieved from HIVDB yielded distinct results between the susceptible and drug-resistant sequences. Most of the resistant sequences show a higher frequency of *major* mutations when compared to the susceptible set. All resistant sequences present at least one mutation in the binding site region, contrasting with 98% of susceptible sequences that do not present any *major* mutations in that site. It is worth noticing that half of the *major* mutations are found in the binding site of resistant sequences. However, when considering the total number of mutations, the increase in the number of mutations per sequence seems to hold a reflection on the increase in the resistance of the observed sequence. Furthermore, binding site *major* mutations are more likely to cause changes on the HIV1-PR binding cleft physicochemical environment when compared with susceptible enzymes which do not have such type of mutations.

A Fast, *Sequence-to-Structure*-, MIF-Based Antiviral Drug Resistance Classifier

The quantification of resistance-conferring mutations in HIV1-PR sequences, using the datasets retrieved from HIVDB, prompted us to further develop a discriminative resistance-classifier approach focused on analysis and comparison of binding-site MIFs. In practice, the proposed workflow involves performing structural modeling of input HIV1-PR sequences using the same template (i.e., 1NH0) and a script (Alves et al., 2018d) that calls *mutate_model.py* (Šali, 2019b) to conduct local energy minimization around the mutated residues of the HIV1-PR structure. Once the generation of structure models is concluded, the modules belonging to the IsoMIF package are deployed for cavity detection (GetCleft module), calculation of MIFs within the selected cavity volume (MIF module), field alignment and quantification of dissimilarities between MIF points computed for the dataset HIV1-PR structural models and those computed for a high quality [Susceptible] reference HIV1-PR structure (1NH0) and, finally, scoring by means of a Tanimoto coefficient (IsoMIF module). The average running time of the workflow is ≈ 77 s per sequence (**Supplementary Figure S1** and **Supplementary Datasheet S1**), considering that this value varies with the amount of mutations present in the HIV1-PR.

TABLE 1 | Tukey's method results to determine outliers.

	Susceptible	*Res_{10}*	*Res_{15}*	*Res_{20}*
Q1	0.0057	0.1075	0.1173	0.0649
Q3	0.0225	0.2041	0.2041	0.2171
IQR	0.0168	0.0966	0.0868	0.1522
Lower Bound	−0.0447*	−0.1823*	−0.1431*	−0.3917*
Upper Bound	0.0729	0.4939	0.4645	0.6737

*Quartile 1 (Q1), Quartile 3 (Q3), Inter Quartile Range (IQR), Upper Bound and Lower Bound values for susceptible sequences dissimilarity coefficient distribution. Upper and Lower Bound were calculated as described in Equation 4, with k = 3. Values above the upper bound and below the lower bound were considered outliers. *Negative values are not realistic lower bounds; the minimum value must be 0.*

Analysis of MIF Dissimilarities in HIV1-PR Binding Site

Figure 4 discloses the frequency of HIV1-PR *sequence-structure* pairs scattered across a spectrum of Tanimoto coefficient (Tc) values (varying from 0.00 to 1.00), in turn reflecting binding-site MIF dissimilarities in the subset of susceptible sequences (containing 7,768 *sequence-structure* pairs) against the selected naïve, template structure. Analyzing this profile of binding site dissimilarities, we observe that there are substantially more susceptible sequences concentrated on lower end of the dissimilarity spectrum. However, a small number of sequences ($N = 81$) present higher values, more visibly around the Tc value of 0.35. Since susceptible HIV1-PR *sequence-structure* pairs display a lower frequency of mutations in the binding site residues, we assume that Tc values deviating from the normal trend may highlight inconsistent data, errors and/or any form of outliers worthy of further investigation.

In order to verify if the higher Tc values could reflect true outliers, Tukey's outlier detection method was used (Tukey, 1949; Hoaglin, 2003). **Table 1** shows the result of applying the statistical Tukey method to the MIF dissimilarity Tc values obtained for the dataset of susceptible *sequence-structure* pairs, and to the [Res_{10}], [Res_{15}], and [Res_{20}] subsets. For each of the four groups, **Figure 5** shows boxplots summarizing the distribution of the MIF dissimilarity Tc values. On the susceptible subset, the higher Tc values were identified as significantly different from the central tendency (values were below the determined lower bound; see Equation 4 in Methods). Looking at the dataset of resistant *sequence-structure* pairs, *extreme* outliers (as described in the Methods section) were only found in the [Res_{10}] subset. These outliers were found to be associated with a software limitation wherein the same reference grid

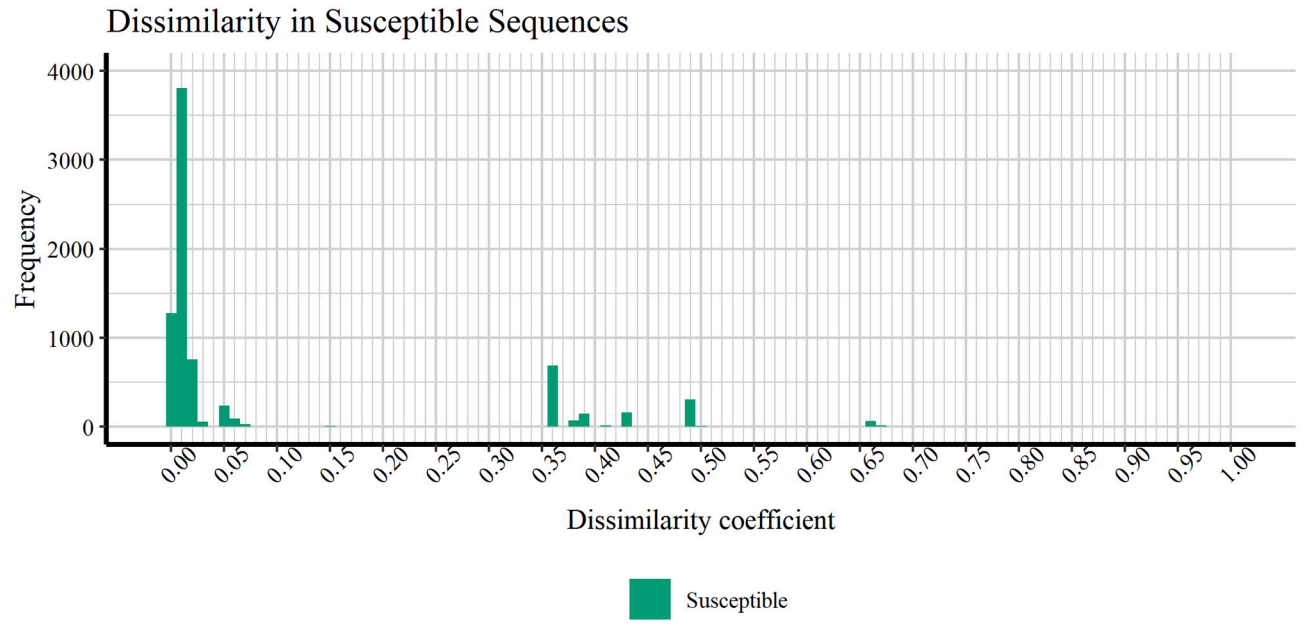

FIGURE 4 | Frequency of susceptible sequences distributed across a range of HIV1-PR binding site 3D dissimilarity coefficients. The maximum value reported is ~0.667.

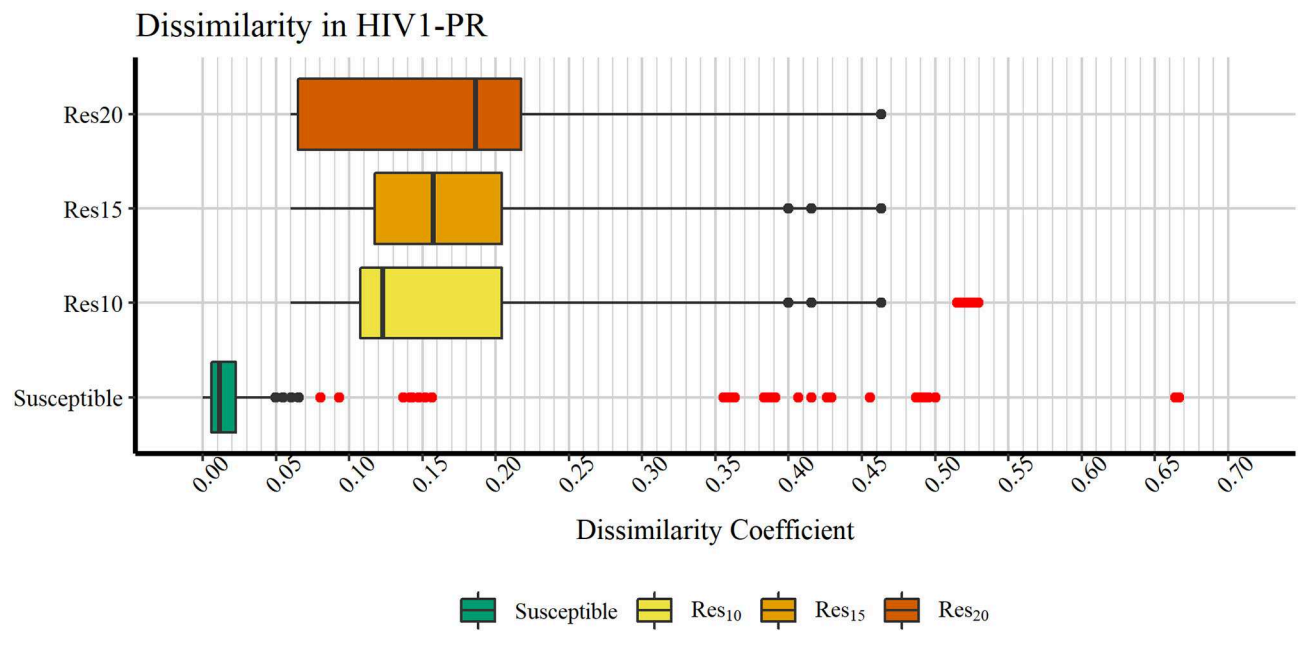

FIGURE 5 | Boxplot representation of dissimilarity coefficients obtained for the HIVDB datasets. The maximum value reported is ~0.667. Outliers are marked in red. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

(generated by GetCleft), covering the entire binding site volume, was not homogeneous across all HIV1-PR structure models. In fact, a wider grid was calculated for some structures when compared to the reference HIV1-PR structure, which resulted on a different number of grid points, consequently leading to an increase of dissimilarities. Thus, these *sequence-structure* pairs were not considered relevant for performance evaluations, as they could introduce performance bias. The Tukey's boxplot analysis thus allowed the identification and removal of *extreme* outliers in the [Susceptible] and [Res_{10}] subsets, resulting in 6269 and 680 HIV1-PR structural models, respectively. The [Res_{15}] and [Res_{20}] subsets remained unchanged with 83 and 60

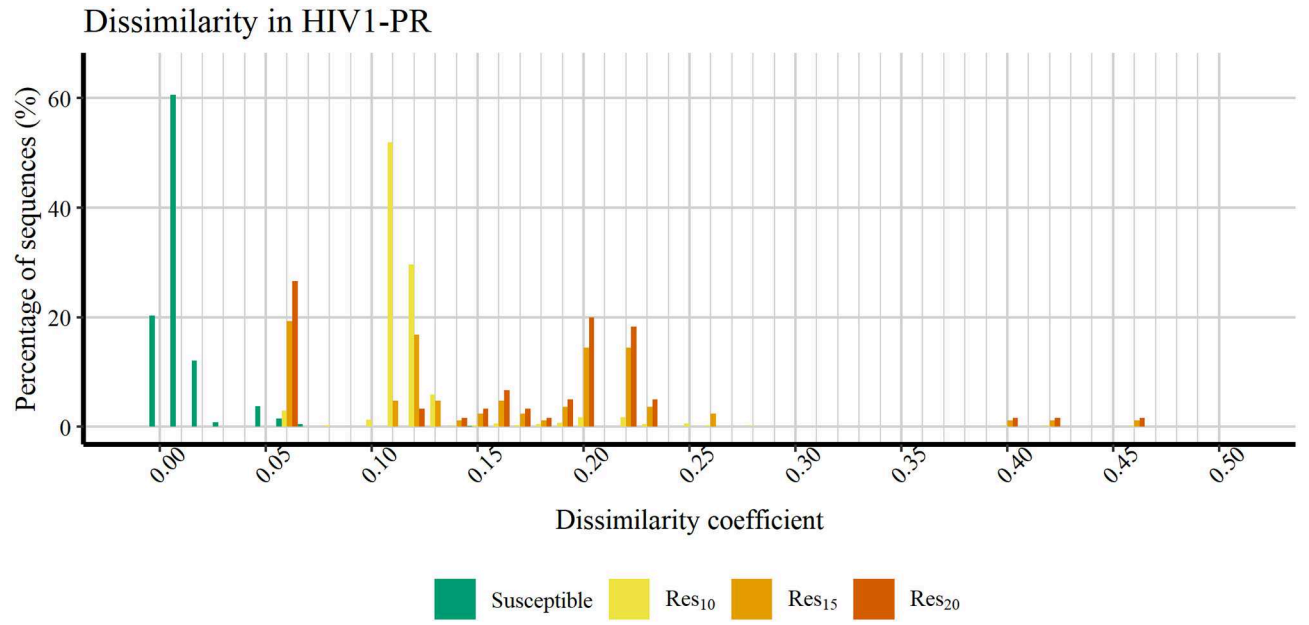

FIGURE 6 | Percentage of sequences distributed across a range of HIV1-PR binding site 3D dissimilarity coefficients. Maximum value reported is ~0.463. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

HIV1-PR structural models, respectively. The resulting dataset has been used for further statistical analysis and as *test set* for performance calculations.

Figure 6 shows a profile of the HIV1-PR binding-site MIF dissimilarities across the susceptible dataset withdrawn of extreme outliers ([Susceptible*]) and the stratified resistant data set (encompassing [Res_{10}], [Res_{15}], and [Res_{20}]) also withdrawn of extreme outliers ([Susceptible*]). As seen, susceptible HIV1-PR structures tend to present very low to null binding-site MIF dissimilarities compared to the ([Susceptible]) structure modeled from the *consensus* sequence. In fact, 93.91% of the *sequence-structure* pairs in the susceptible group show dissimilarities lower than 0.02, indicating a considerable degree of conservation within the binding site. Overall, these results show a segregation between susceptible and resistant *sequence-structure* pairs, when analyzing their binding-site MIF dissimilarities against a susceptible reference sequence-structure pair, suggesting that our method is able to quantitatively capture differences among susceptible and resistant HIV1-PR structures.

Evaluation of the Classification Performance of Our Drug Resistance Classifier

At the current stage of development, the proposed workflow only performs binary classification, meaning that each input sequence gets classified as either susceptible or resistant. Sequence data are used exclusively for the generation of the structural models on which dissimilarities are analyzed, but not to aid the classification itself. It is worth highlighting that our workflow relies on the detection of structural and chemical changes in viral enzymes that dictate susceptibility or resistance to drugs—rather than on the training of predictive models using sequences with known phenotypic response to drugs. Therefore, instead of using performance evaluation methods, such as cross-validation, that assess the impact of hiding a portion of training data (observations) on the accuracy of the resulting predictions, we resorted to the calculation of metrics of overall performance of our binary classifier.

The Receiver Operating Characteristic (ROC) curve was used to assess the overall discriminatory performance of our method. The score assigned to each dataset entry (here used for testing), corresponding to binding-site dissimilarities between each input *sequence-structure* pair and the template *consensus sequence-structure*, were thus plotted as a ROC curve. ROC curves are conceptually simple plots that depicts a binary classifier's discriminative capability as its discrimination threshold is varied. Such graphical plots are created by plotting the method's true positive rate (sensitivity) against its false positive rate (1-specificity), at varying thresholds. The area under the ROC curve (ROC AUC) value is a single scalar value varying between 0 and 1, providing a measure of the overall discriminatory power of the method. A ROC AUC value of 1 (or 100%) entails a *perfect* discrimination, a value of 0.5 represents random classification, while values above 0.8 are commonly accepted as indicators of an acceptable discriminatory performance (Fawcett, 2006; Pines and Everett, 2008; Powers, 2011; Tape). Furthermore, several performance measures, such as the Sensitivity (Equation 5), Specificity (Equation 6), Accuracy (Equation 7), and MCC (Equation 8) were also determined.

Figure 7 represents the obtained ROC curves and their respective ROC AUC values for the susceptible and resistant HIV1-PR binding-site MIF dissimilarities. ROC AUC values for [Res_{10}], [Res_{15}], and [Res_{20}] subsets were

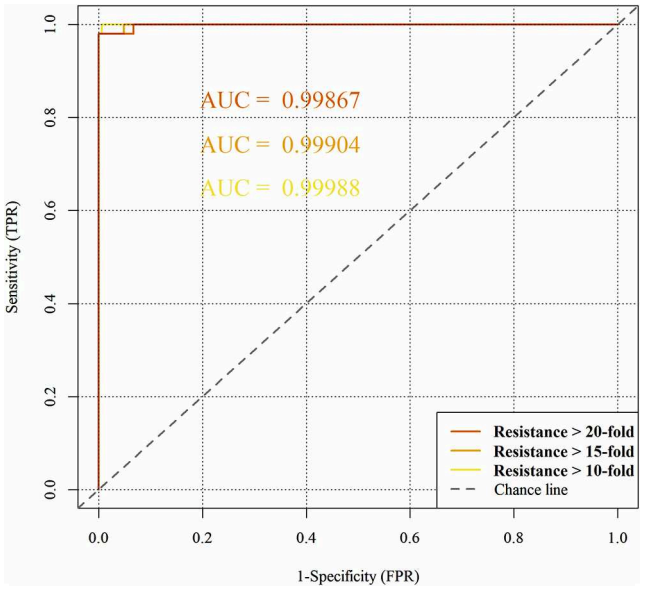

FIGURE 7 | Predictive performance of the binary classification (resistant vs. susceptible) produced by the algorithm/workflow presented herein, quantified by means of Receiver Operating Characteristic (ROC) curves and their respective Area Under the Curve (ROC AUC). The colors represent the ROC curves as follows: yellow for HIV1-PR sequences associated with 10-fold resistance; orange for HIV1-PR sequences associated with 15-fold resistance; and dark orange for HIV1-PR sequences associated with 20-fold resistance.

found to be similarly very high-—0.9999, 0.9990, and 0.9987, respectively – suggesting that the method holds significant discriminatory power to distinguishing susceptible from fully resistant HIV1-PR *sequence-structure* pairs—based on their binding-site MIF dissimilarities to the [Susceptible] reference *sequence-structure* pair.

We have also used ROC curve analysis to guide the definition of an *optimal* discrimination threshold based on Youden's index (Equation 9) (Youden, 1950). The optimal threshold observed corresponded to a 0.06 dissimilarity Tc for all [Res_{10}], [Res_{15}], and [Res_{20}] subsets. **Table 2** presents the values of each performance metric obtained for each subset, when applying a classification threshold of 0.06. At this classification cut-off, the specificities and sensitivities were found to be 0.997 and 0.994 for the [Res_{10}] subset, 0.997 and 0.952 for the [Res_{15}] subset and 0.997 and 0.933 for the [Res_{20}] subset, respectively. In all cases, there is strong discriminative performance toward susceptibility or resistance—as it can be appreciated by the high accuracy values highlighted in **Table 2**. Nevertheless, the best results are found for the [Res_{10}] subset, with an accuracy of about 0.997. On the other hand, the subsets with increasing degree of resistance, [Res_{15}] and [Res_{20}], show only slightly worst results concerning Sensitivity determined at a threshold of 0.06.

The overall predictive performance of our method was also evaluated by the Matthews correlation coefficient (MCC) on the three resistant subsets, which summarizes the sensitivity and the specificity of a classification method within a unique value, also varying between 0 and 1. A higher value of MCC indicates that the method has a better discriminatory performance. For the [Res_{10}], [Res_{15}], and [Res_{20}] groups, MCC values of 0.982, 0.874, and 0.833 were, respectively, obtained. Still, such performance metrics seems to highlight the clear potential of our MIF-based method to predict drug resistance, especially within the most populated [Res_{10}] group (MCC value close to 1).

TABLE 2 | Performance metrics obtained using a dissimilarity threshold of 0.0603.

Dissimilarity Threshold = 0.0603	Res_{10}	Res_{15}	Res_{20}
ROC AUC	0.99988	0.99904	0.99867
Sensitivity	0.994118	0. 951807	0. 933333
Specificity	0.992184	0. 992184	0. 992184
Accuracy	0.99669	0.996379	0.996366
MCC	0.98151	0.874199	0.833085

Positioning and Differentiation vs. Sequence-Based, PI-Resistance Prediction Tools

More than a decade ago, Lengauer and Sing pointed out the lack of commonly agreed benchmark (or test) datasets to assess and compare the performance of different prediction methods (Lengauer and Sing, 2006). The amount of available information on matched HIV genotype–resistance phenotype has increased significantly over recent years, with HIVDB embodying an important role as a centralized data repository (Rhee et al., 2003). As expected, sequence-based methods can make use of as much information as available to train their predictions, resulting in that they become proficient at "predicting" the phenotypic response for the sequences they have been trained on. Only in a few cases do we witness a concern in drawing prospective validation on unseen sequence sets and in making those test sets available to the community (Tarasova et al., 2018). This hinders the design of fair comparisons with methods that do not make direct use of sequence data for training, such as the one we propose here. On the other hand, over the past years genotypic-based methods have reached a level of sophistication that allows them to perform resistance predictions to specific drugs, exclusively based on sequence data matched to phenotypic response, while, at its current stage of development, our MIF-based method can only perform binary classification (susceptible or resistant) of input sequences.

Taken together, these aspects render the comparison of our algorithm with existing, sequence-trained, multi-classification predictors *non-trivial* to say the least. Further developments of our methodology, aiming at a more exhaustive exploration of specific MIF areas around the mutated binding sites, may enable stratification of classification into multiple drug classes by detecting the determinants of resistance to specific PIs. For the time being, we center the analysis of differentiation of our method on the answer to a recurrent question in the mind virologists or physicians who prescribe HIV-1 medications: *would it be possible to accurately predict whether a new, unknown HIV-1 strain will be susceptible to known PIs?*

TABLE 3 | Performance metrics for exemplary sequence-based prediction tools tested against the datasets compiled in this work.

PI-resistance predictor	Sensitivity$_{(A)}$	Sensitivity$_{(B)}$	Specificity	FN$_{(A)}$‡	FN$_{(B)}$‡	FP‡
HIV-GRADE 07/2019	1.0000	0.1471	0.9809	0	580 *12*	120 *8*
ANRS 29_11/2018	1.0000	0.1868	0.8493	0	553 *14*	945 *61*
HIVdb 8.9.1	1.0000	1.0000	0.8818	0	0	741 *21*
Rega 10.0.0	1.0000	0.0838	0.9804	0	623 *7*	123 *10*
MIF-based Drug Resistance Classifier†	0.9941	0.9941	0.9922	4 *1*	4 *1*	49 *11*

†*The proposed MIF-based drug resistance classifier is shown in the last row for comparison purposes.*

‡*False negatives (FN) corresponds to the number of sequences belonging to the Resistant* dataset (withdrawn of extreme outliers) that were predicted susceptible to all PIs. False positives (FP) corresponds to the number of sequences belonging to the Susceptible* dataset (withdrawn of extreme outliers) that were predicted resistant to at least one PI. In italics are indicated the number of viral isolates to which the sequences misclassified as FP belong. Rules for sensitivity analysis in (1) benchmark A [Sensitivity$_{(A)}$]: resistance to one or more PIs is considered a correct prediction; and (2) benchmark B [Sensitivity$_{(B)}$]: resistance to all PIs is considered a correct prediction.*

In order to answer to this question, we first converted our *test set* containing susceptible and resistant HIV1-PR sequences withdrawn of *extreme* outliers ($N = 6{,}269$ [[Susceptible*]] and $N = 680$ [[Resistant*]], respectively) into codon code, using the EMBOSS Backtranseq online tool (Madeira et al., 2019a,b), and then submitted it to the HIV-GRADE web server (Obermeier et al., 2012a,b) for comparison with the sequence-based algorithms ANRS-rules (Brun-Vézinet et al., 2003), HIVdb (Rhee et al., 2003; Tang et al., 2012) and Rega (Van Laethem et al., 2002; Camacho et al., 2017). Unexpectedly, we were not able to obtain predictions from *geno2pheno* via HIV Grade due to a technical issue of the web platform. To eschew this problem, we tried to submit the *test set* directly through *geno2pheno*'s web server, but the interface is limited to an unpractical maximum of 20 sequences per run.

Because the existing sequence-based interpretation systems try to predict phenotypical susceptibility or resistance to the individual drugs for a given genotype, whereas our approach only performs binary classification (susceptibility or resistance to all PIs), in order to draw comparison between the methods we tried to "level the playing field" by converting the predictions made by *sequence-based* algorithms into simpler binary classifications. In a first benchmark (benchmark A), the prediction outputs were converted into (i) susceptibility to all PIs ([Susceptible]) or (ii) resistance to any PI (*Resistant*). In a second, more challenging benchmark (benchmark B), the outputs were encoded as either (i) susceptible to all PIs (*Susceptible*) or (ii) resistant to all PIs (*Resistant*). The full list of criteria applied to the conversion of multiple classifiers into binary classification is given in **Supplementary Table S2**. The full raw output of HIV-GRADE is available in **Supplementary Datasheet S2**.

The ability to accurately predict the susceptibility of the input sequences to all PIs was assessed by determining the rate of correct predictions, with reflection into the calculated methods' Sensitivity (Equation 5) and Specificity (Equation 6). **Table 3** lists calculated performance metrics for the *sequence-based* algorithms on both benchmarks A and B, contrasted with the performance of our *sequence-to-structure*-, MIF-based algorithm. Sensitivity$_{(A)}$ and the number of detected false negatives FN$_{(A)}$ translate the methods' ability to classifying a HIV1-PR sequence known to be resistant to all PIs as *Resistant to at least one PI*. In contrast, Sensitivity$_{(B)}$ and FN$_{(B)}$ translate the methods' ability to correctly predict the same sequences (known to be resistant to all PIs) as *resistant to all PIs*. From the methods' sensitivity viewpoint, the assessment of the results of both benchmarks A and B has been important to counterbalance the crudeness of the conversion of a multiple classifier of resistance toward specific PIs into a binary classification. Benchmark A clearly biases sensitivity in favor of a multi-classifier by considering any resistance prediction (in number or kind of PI) for sequences known to be resistant to all PIs as *correct*, whereas benchmark B offers a more stringent evaluation of sensitivity wherein only *resistant-to-all-PIs* predictions for the same set of fully resistant sequences are considered as *correct*.

As expected the discriminatory power of the methods in benchmark A is in stark contrast with that calculated for benchmark B. Sensitivity$_{(A)}$ suggests that *sequence-based* methods slightly outperform our *sequence-to-structure*-, MIF-based classifier, with 100% correct predictions of *Resistant* sequences vs. a Sensitivity$_{(A)}$ value of 0.994 obtained by our method. By contrast, benchmark B shows a considerable drop in performance by *sequence-based* methods at correctly predicting HIV1-PR sequences resistant to all PIs—aside HIVdb, which retains a Sensitivity of 1.000.

The results in **Table 3** indicate that our workflow outperforms all other algorithms at identifying sequences susceptible to all PIs, with a Specificity of approximately 0.992, while its sequence-based counterparts display Specificities ranging from approximately 0.849 to 0.981. Still, it is worth noting that the large number of FP from the other *sequence-based* methods mostly come from the same isolates, similarly as mentioned above for FN$_{(B)}$. This fact highlights the advantage of accounting for structural information besides genotypic data. While MIFs allow searching for differences in the structural and physicochemical

environment of proteins, which might not be significantly affected by mutations for similar amino acids, *sequence-based* approaches will consistently search for mutations at positions of interest and consistently assign them the same classification. At an early-stage of development, our workflow's performance is quite satisfactory, considering that the ability of correctly classifying a sequence as susceptible to all PIs is a highly relevant step at the beginning of antiretroviral therapy—where a false positive weights more on the flexibility of first-line therapy regimens and, consequently, quality of life of the patient.

CONCLUSION AND FUTURE PERSPECTIVES

In recent years, the availability of data in the form of matched HIV genotype–resistance phenotype has expanded greatly, enabling further training of statistical learning methods relating genotype to different levels of phenotypic resistance and against specific drugs. However, in spite of the increased access to and routine sequencing of HIV's genome in many countries, as well as the constant evolution of machine learning (ML)-based techniques, HIV's high mutation rate (estimated in 3×10^{-5} per nucleotide per replication) will continue to pose significant challenges: not only in terms of the constant demand for curation of genotypic and phenotypic data to be fed into ML algorithms, but also from the viewpoint of the interpretability and translation of said data into knowledge to assist the design of novel anti-microbial agents. Therefore, the exploration of innovative structure-based *in silico* approaches to the prediction of drug resistance, focusing at the molecular interface that bridges to drug design, holds clear interest and appeal as alternative or complement to some of the most developed sequence-based statistical methods.

In this contribution, we propose a novel approach to drug resistance prediction, which captures structural and physicochemical modifications induced by mutations in the binding site of an extensively studied viral target, HIV1-PR. We demonstrate that, even at an early, proof-of-principle stage of development, our methodology can identify HIV1-PR *sequence-structure* pairs belonging to three levels of increasing resistance—with impressively high accuracy—thus anticipating, on a purely structural basis, whether a given HIV1-PR sequence will translate into phenotypic resistance or susceptibility to PIs. Since our *sequence-to-structure*-based classifier does not rely on *training* from genotypic data and only uses an individual input sequence to derive the corresponding viral enzyme structure and yield a prediction, its potential real-world value in supporting clinical decision is clearly relevant. Due to the fact that the proposed workflow produced predictions of complete drug susceptibility to the HIV1-PR datasets with high predictive accuracy, said results highlight this methodology as a potential valuable resource on clinical practice. Being able to use the clinical isolate sequence data to accurately predict susceptibility to known PIs, before starting a therapeutic regimen, is of paramount importance to allow the initiation of PI-based therapy with the less expensive 1st generation PIs, resulting in an economic benefit to the healthcare systems. Importantly, even though the method performs analysis on thousands of structural data points (atomic coordinates and MIF points), classification into *susceptible* or *resistant* takes place in a *couple-of-minutes* time scale.

It is worth emphasizing, nevertheless, that there is obvious room for methodological improvement and expansion. The upgrade to multi-classification functionality, where target structures known to be susceptible to specific inhibitors and drugs are used as template for structural modeling, is a critical milestone that will pave the way to predicting resistance to those specific anti-microbial agents. The growing amount of three-dimensional structural data on microbial target-inhibitor complexes, coupled with more elaborate use of sequence data, fuels our belief in that an improved *sequence-to-structure* -, MIF-based drug resistance classifier, will be able to combine the strengths and overcome the shortcomings of current approaches.

Claims of greatness must be backed by adequate validation designs. While the current version of our workflow does not allow drawing comprehensive and direct comparisons with more advanced sequence-based predictors of resistance to specific HIV1-PR inhibitors, further developments to our method will also be accompanied by the assembly and sharing of stratified benchmark sets of susceptible and resistant microbial target sequences—enabling fairer comparisons to be made both by ourselves and the scientific community.

As implied in our concluding words, a clear expectation around this work involves extending the application of our method to other targets, other than HIV1-PR, with inherent and multiple patterns of genetic variation. We realize, however, that this expectation may only be fulfilled if workable amounts of data are shared among the scientific community. Undoubtedly, one of the most critical aspects facing drug resistance prediction is the development of community-wide efforts to prepare and share useful datasets and tools to facilitate improvement and performance evaluation of existing and novel methodologies—which should be a clear priority for researchers working in the field. By basing its development on the use of freeware, our method is freely-available for non-commercial use.

To conclude, we see the results presented here as a promising example of the potential application of combined sequence- and structure-based *in silico* methods to achieve a more detailed interpretation and prediction of the impact of mutations in drug resistance. The ever-increasing emergence and widespread of drug-resistance calls in for the development of more efficient strategies to combat microbial threats in several fronts—be that in the drug discovery research setting or the clinical and medical therapeutic decision realm.

DATA AVAILABILITY STATEMENT

The datasets analyzed and scripts for this study can be found in the PI-resistance_Prediction GitHub [https://github.com/subject-am/PI-resistance_Prediction]. Raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

CS and RB developed the idea for the present work and provided critical revisions. NA, AM, and JL contributed equally to its conception, literature search, and manuscript writing. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

NA, AM, JL, CS, and RB thank Daniela Vaz, João Vaz, and Vítor Duque for the informative discussions on the HIV1 drug resistance topic, which lead to devising and developing of the present work. NA, AM, JL, CS, and RB also thank the Coimbra Chemistry Centre (CQC) supported by the Portuguese Agency for Scientific Research, Foundation for Science and Technology (FCT), through Project UID/QUI/00313/2019. NA, AM, and JL acknowledge the MedChemTrain Ph.D. programme (PD/00147/2013) in Medicinal Chemistry—Ministry of Science, Technology, and Higher Education (MCTES), Portugal—for Ph.D. fellowships PD/BD/135287/2017, PD/BD/135289/2017 and PD/BD/135292/2017, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00243/full#supplementary-material

REFERENCES

Agniswamy, J., Louis, J. M., Roche, J., Harrison, R. W., and Weber, I. T. (2016). Structural studies of a rationally selected multi-drug resistant HIV-1 protease reveal synergistic effect of distal mutations on flap dynamics. *PLoS ONE* 11:e0168616. doi: 10.1371/journal.pone.0168616

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Alves, N. G., Mata, A. I., and Luís, J. P. (2018a). *GitHub: FilterMajor.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/filterMajor.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018b). *GitHub: FilterMinor.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/filterMinor.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018c). *GitHub: MutModels.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/model-structures/MutModels.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018d). *GitHub: Pattern_HIVp.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/model-structures/pattern-HIVp.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018e). *GitHub: Ref_process.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/isomif_run/ref_process.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018f). *GitHub: Separate_sets.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Separate_sets.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2019a). *GitHub: Count_mut.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/count_mut.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2019b). *GitHub: HIV1predict.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/master/bin/HIV1predict.sh (accessed November 30, 2019).

Artese, A., Cross, S., Costa, G., Distinto, S., Parrotta, L., Alcaro, S., et al. (2013). Molecular interaction fields in drug discovery: recent advances and future perspectives. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 594–613. doi: 10.1002/wcms.1150

Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., et al. (2003). Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.* 31, 3850–3855. doi: 10.1093/nar/gkg575

Bonet, I. (2015). Machine learning for prediction of HIV drug resistance: a review. *Curr. Bioinform.* 10, 579–585. doi: 10.2174/1574893610666151008011731

Bron, C., and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16, 575–577. doi: 10.1145/362342.362367

Brun-Vézinet, F., Descamps, D., Ruffault, A., Masquelier, B., Calvez, V., Peytavin, G., et al. (2003). Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS* 17, 1795–1802. doi: 10.1097/00002030-200308150-00008

Camacho, R., Laethem, K., Van Geretti, A. M., Verheyen, J., Paredes, R., and Vandamme, A.-M. (2017). *Algorithm for the Use of Genotypic HIV-1 Resistance Data (Version Rega v10.0.0)*. Available online at: https://rega.kuleuven.be/cev/avd/software/rega-hiv1-rules-v10.pdf

Cao, Z. W., Han, L. Y., Zheng, C. J., Ji, Z. L., Chen, X., Lin, H. H., et al. (2005). Computer prediction of drug resistance mutations in proteins. *Drug Discov. Today* 10, 521–529. doi: 10.1016/S1359-6446(05)03377-5

Chartier, M., and Najmanovich, R. (2015). Detection of binding site molecular interaction field similarities. *J. Chem. Inf. Model* 55, 1600–1615. doi: 10.1021/acs.jcim.5b00333

Cruciani, G. (2005). *Molecular Interaction Fields*. Weinheim: FRG, Wiley-VCH Verlag GmbH and Co. KGaA.

draw.io. (2005). *Flowchart Maker & Online Diagram Software*. Available online at: https://www.draw.io/ (accessed October 25, 2019).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin. Biochem. Rev.* 29(Suppl 1), S83–S87.

Gaudreault, F., Morency, L.-P., and Najmanovich, R. J. (2015). NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* 31, 3856–3858. doi: 10.1093/bioinformatics/btv458

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28, 849–857. doi: 10.1021/jm00145a002

Hao, G., Yang, G., and Zhan, C. (2010). Computational mutation scanning and drug resistance mechanisms of HIV-1 protease inhibitors. *J. Phys. Chem. B* 114, 9663–9676. doi: 10.1021/jp102546s

Hao, G.-F., Guang-Fu, Y., and Zhan, C.-G. (2012). Structure-based methods for predicting target mutation-induced drug resistance and rational drug design to overcome the problem. *Drug Discov. Today* 17, 1121–1126. doi: 10.1016/j.drudis.2012.06.018

Hoaglin, D. C. (2003). John W. Tukey and data analysis. *Stat. Sci.* 18, 311–18. doi: 10.1214/ss/1076102418

Hou, T., and Yu, R. (2007). Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. *J. Med. Chem.* 50, 1177–1188. doi: 10.1021/jm0609162

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Irwin, K. K., Renzette, N., Kowalik, T. F., and Jensen, J. D. (2016). Antiviral drug resistance as an adaptive process. *Virus Evol.* 2:vew014. doi: 10.1093/ve/vew014

Jenwitheesuk, E., and Samudrala, R. (2005). Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir. Ther.* 10, 157–166.

Khalid, Z., and Sezerman, O. U. (2018). Prediction of HIV drug resistance by combining sequence and structural properties. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 966–973. doi: 10.1109/TCBB.2016.2638821

Khan, R. A., and Brandenburger, T. (2019). *ROCit: Performance Assessment of Binary Classifier with Visualization*. Available online at: https://cran.r-project.org/package=ROCit

King, N. M., Prabu-Jeyabalan, M., Nalivaika, E. A., Wigerinck, P., de Bethune, M.-P., and Schiffer, C. A. (2004). Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.* 78, 12012–12021. doi: 10.1128/JVI.78.21.12012-12021.2004

Lengauer, T., and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.* 4, 790–797. doi: 10.1038/nrmicro1477

Madeira, F., Mi Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019a). EMBOSS Backtranseq. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz26

Madeira, F., Mi Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019b). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

Martínez, L., Andreani, R., and Martínez, J. M. (2007). Convergent algorithms for protein structural alignment. *BMC Bioinformatics* 8:306. doi: 10.1186/1471-2105-8-306

Mason, S., Devincenzo, J. P., Toovey, S., Wu, J. Z., and Whitley, R. J. (2018). Comparison of antiviral resistance across acute and chronic viral infections. *Antiviral Res.* 158, 103–112. doi: 10.1016/j.antiviral.2018.07.020

Masso, M., and Vaisman, I. I. (2013). Sequence and structure based models of HIV-1 protease and reverse transcriptase drug resistance. *BMC Genomics* 14:S3. doi: 10.1186/1471-2164-14-S4-S3

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

McKeegan, K. S., Borges-Walmsley, M. I., and Walmsley, A. R. (2002). Microbial and viral drug resistance mechanisms. *Trends Microbiol.* 10, 8–14. doi: 10.1016/S0966-842X(02)02429-0

Nayak, C., Chandra, I., and Singh, S. K. (2019). An *in silico* pharmacological approach toward the discovery of potent inhibitors to combat drug resistance HIV-1 protease variants. *J. Cell. Biochem.* 120, 9063–9081. doi: 10.1002/jcb.28181

Obermeier, M., Pironti, A., Berg, T., Braun, P., Däumer, M., Eberle, J., et al. (2012a). HIV-GRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology* 55, 102–107. doi: 10.1159/000331999

Obermeier, M., Pironti, A., Berg, T., Braun, P., Däumer, M., Eberle, J., et al. (2012b). HIV-GRADE: HIV-1 Tool. *Intervirology* 55, 102–107.

Pawar, S., Wang, Y.-F., Wong-Sam, A., Agniswamy, J., Ghosh, A. K., Harrison, R. W., et al. (2019). Structural studies of antiviral inhibitor with HIV-1 protease bearing drug resistant substitutions of V32I, I47V and V82I. *Biochem. Biophys. Res. Commun.* 514, 974–978. doi: 10.1016/j.bbrc.2019.05.064

Pines, J. M., and Everett, W. W. (2008). *Evidence-Based Emergency Care*, 2nd Edn, eds J. M. Pines, C. R. Carpentee, A. S. Raja, and J. D. Schuur. Oxford: Blackwell Publishing Ltd.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* 2, 37–63. Available online at: http://bioinfopublication.org/viewhtml.php?artid=BIA0001114

Qiu, X., and Liu, Z.-P. (2011). Recent developments of peptidomimetic HIV-1 protease inhibitors. *Curr. Med. Chem.* 18, 4513–4537. doi: 10.2174/092986711797287566

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Found. Stat. Comput. Available online at: https://www.r-project.org/

RCSB PDB (2000). *RCSB PDB*. Available at: https://www.rcsb.org/ (accessed September 7, 2019).

[Res_{10}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 10-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res10

[Res_{15}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 15-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res15

[Res_{20}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 20-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res20

[Resistant*] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 10-fold resistance to all inhibitors, without extreme outliers. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Resistant_noOut

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. U. S. A.* 103, 17355–17360. doi: 10.1073/pnas.0607274103

Riemenschneider, M., Hummel, T., and Heider, D. (2016). SHIVA - a web application for drug resistance and tropism testing in HIV. *BMC Bioinformatics* 17:314. doi: 10.1186/s12859-016-1179-2

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., and Schisterman, E. F. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical J.* 50, 419–430. doi: 10.1002/bimj.200710415

Šali, A. (2019a). *Modeller*. Available at: https://salilab.org/modeller/ (accessed August 20, 2019).

Šali, A. (2019b). *Modeller Wiki*. Available online at: https://salilab.org/modeller/wiki/Mutate model (accessed September 30, 2019).

Šali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815. doi: 10.1006/jmbi.1993.1626

Sheik Amamuddy, O., Bishop, N. T., and Tastan Bishop, Ö. (2018). Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. *Sci. Rep.* 8:17938. doi: 10.1038/s41598-018-36041-8

Stanford University (1998a). *HIVDB: Consensus B Amino Acid Sequences*. Available online at: https://hivdb.stanford.edu/pages/documentPage/consensus_amino_acid_sequences.html (accessed November 2, 2019).

Stanford University (1998b). *HIVDB: Genotype-Phenotype Datasets*. Available online at: https://hivdb.stanford.edu/pages/genopheno.dataset.html (accessed November 2, 2018).

Stanford University (1998c). *HIVDB: HIV Drug Resistance Database*. Available online at: https://hivdb.stanford.edu/ (accessed October 23, 2019).

Stanford University (1998d). *HIVDB: PI Resistance Notes*. Available online at: https://hivdb.stanford.edu/dr-summary/resistance-notes/PI/ (accessed October 23, 2019).

Strasfeld, L., and Chou, S. (2010). Antiviral drug resistance: mechanisms and clinical implications. *Infect. Dis. Clin. North Am.* 24, 413–437. doi: 10.1016/j.idc.2010.01.001

[Susceptible] Alves, Nuno G., Ana Isabel Mata and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University susceptible to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-susceptible

[Susceptible*] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University susceptible to all inhibitors, without *extreme* outliers. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Susceptible_noOut

Tang, M. W., Liu, T. F., and Shafer, R. W. (2012). The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology* 55, 98–101. doi: 10.1159/000331998

Tape, T. G. (1990). *Interpreting Diagnostic Tests*. University of Nebraska Medical Center. Available online at: http://gim.unmc.edu/dxtests/ (accessed October 24, 2019).

Tarasova, O., Biziukova, N., Filimonov, D., and Poroikov, V. (2018). A computational approach for the prediction of HIV resistance based on amino acid and nucleotide descriptors. *Molecules* 23:2751. doi: 10.3390/molecules23112751

Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P., and Arora, S. K. (2011). Prediction of drug-resistance in HIV-1 subtype C based on protease sequences

from ART naive and first-line treatment failures in North India using genotypic and docking analysis. *Antiviral Res.* 92, 213–218. doi: 10.1016/j.antiviral.2011.08.005

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics* 5:99. doi: 10.2307/3001913

Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., Perna, C. F., and Vandamme, A.-M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir. Ther.* 7, 123–9.

Vercauteren, J., and Vandamme, A.-M. (2006). Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Res.* 71, 335–342. doi: 10.1016/j.antiviral.2006.05.003

Vere Hodge, A., and Field, H. J. (2011). "General mechanisms of antiviral resistance," in *Genetics and Evolution of Infectious Disease*, ed M. Tibayrenc (Amsterdam: Elsevier), 339–362.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., et al. (2019). *gplots: Various R Programming Tools for Plotting Data*. Available online at: https://cran.r-project.org/package=gplots

Weber, I. T., and Agniswamy, J. (2009). HIV-1 protease: structural perspectives on drug resistance. *Viruses* 1, 1110–1136. doi: 10.3390/v1031110

Weber, I. T., and Harrison, R. W. (2016). Tackling the problem of HIV drug resistance. *Postepy Biochem.* 62, 273–279.

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Günthard, H. F., Paredes, R., et al. (2019). 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 27, 111–121. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31634862

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. Available online at: https://ggplot2-book.org/. doi: 10.1007/978-0-387-98141-3

Wlodawer, A., and Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* 62, 543–585. doi: 10.1146/annurev.bi.62.070193.002551

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285, 1735–1747. doi: 10.1006/jmbi.1998.2401

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32–35.

Yu, X., Weber, I. T., and Harrison, R. W. (2014). Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. *BMC Genomics* 15:S1. doi: 10.1186/1471-2164-15-S5-S1

Zhang, J., Rhee, S.-Y., Taylor, J., and Shafer, R. W. (2005). Comparison of the precision and sensitivity of the antivirogram and phenosense HIV drug susceptibility assays. *J. Acquir. Immune Defic. Syndr.* 38, 439–444. doi: 10.1097/01.qai.0000147526.64863.53

Conflict of Interest: CS and RB are cofounders of the company BSIM Therapeutics, however all work reported in this article was carried out at the University of Coimbra.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.