biomolecules

MDPI

Article

AI-Aided Search for New HIV-1 Protease Ligands

Roberto Arrigoni [1], Luigi Santacroce [2], Andrea Ballini [3,*] and Luigi Leonardo Palese [4,*]

[1] Bioenergetics and Molecular Biotechnologies (IBIOM), CNR Institute of Biomembranes, 70125 Bari, Italy; r.arrigoni@ibiom.cnr.it
[2] Interdisciplinary Department of Medicine (DIM), University of Bari Aldo Moro, 70124 Bari, Italy; luigi.santacroce@uniba.it
[3] Department of Clinical and Experimental Medicine, University of Foggia, 71122 Foggia, Italy
[4] Department of Translational Biomedicine and Neurosciences—(DiBraiN), University of Bari Aldo Moro, 70124 Bari, Italy
* Correspondence: andrea.ballini@me.com (A.B.); luigileonardo.palese@uniba.it (L.L.P.)

Abstract: The availability of drugs capable of blocking the replication of microorganisms has been one of the greatest triumphs in the history of medicine, but the emergence of an ever-increasing number of resistant strains poses a serious problem for the treatment of infectious diseases. The search for new potential ligands for proteins involved in the life cycle of pathogens is, therefore, an extremely important research field today. In this work, we have considered the HIV-1 protease, one of the main targets for AIDS therapy. Several drugs are used today in clinical practice whose mechanism of action is based on the inhibition of this enzyme, but after years of use, even these molecules are beginning to be interested by resistance phenomena. We used a simple artificial intelligence system for the initial screening of a data set of potential ligands. These results were validated by docking and molecular dynamics, leading to the identification of a potential new ligand of the enzyme which does not belong to any known class of HIV-1 protease inhibitors. The computational protocol used in this work is simple and does not require large computational power. Furthermore, the availability of a large number of structural information on viral proteins and the presence of numerous experimental data on their ligands, with which it is possible to compare the results obtained with computational methods, make this research field the ideal terrain for the application of these new computational techniques.

Keywords: HIV-1 protease; HIV protease inhibitors; molecular docking; drug resistance; artificial intelligence; autoencoder

Citation: Arrigoni, R.; Santacroce, L.; Ballini, A.; Palese, L.L. AI-Aided Search for New HIV-1 Protease Ligands. *Biomolecules* **2023**, *13*, 858. https://doi.org/10.3390/biom13050858

Academic Editors: Umesh Desai and Daniel Afosah

Received: 24 April 2023
Revised: 15 May 2023
Accepted: 16 May 2023
Published: 18 May 2023

1. Introduction

In the early 1980s, a new viral infection was recognized, with a pandemic trend that is still ongoing. Acquired immunodeficiency syndrome (AIDS), which has had dramatic clinical implications for many years, is caused by a retrovirus known as human immunodeficiency virus type 1 (HIV-1). It is estimated that this virus has caused around 40 million deaths worldwide, and currently, there are about 38 million patients infected by HIV-1 [1]. Despite significant progress in understanding the immune response to the virus and developing vaccines [2–4], the disease is now controlled only by antiviral drugs [5,6]. These molecules affect three viral enzymes: protease, reverse transcriptase, and integrase. The availability of these drugs has radically changed the prognosis and quality of life of HIV-infected patients, so much that the reassuring certainty that HIV-1 infection has become a chronic, but controllable, condition has spread in public opinion. However, this optimistic vision soon collided with the emergence of strains resistant to one or more drugs [7–10]. The rate of infections with resistant strains has increased significantly in recent years, especially in North America and sub-Saharan Africa, to the point of becoming a real threat to public health.

HIV-1 protease, encoded as part of the Gag-Pol polyprotein in the viral genome, is responsible for the maturation of the Gag-Pol and Gag precursors and has an essential role in the viral replication cycle [11]. For this reason, it was quickly realized that molecules capable of blocking this enzyme were excellent drug candidates for the treatment of AIDS [11,12]. This enzyme is a C2 symmetric homodimer of subunits containing 99 residues [13,14]. Its active site is located at the dimer interface and contains a catalytic aspartate (D25) in the sequence signature DTG. Three regions can be recognized in each monomer: one involved in the enzyme dimerization (residues 1–4 and 95–99); the core region (residues 10–32 and 63–85 of each monomer), which participates to the catalytic site, as well as dimerization; and the flap region consisting of two solvent exposed loops (residues 33–43 of each chain) and two flexible, glycine-rich β-hairpins (residues 44–62). The flexible flaps cap the catalytic triad, and upon substrate- or inhibitor-binding, these plug the active site. The error-prone replication of the HIV-1 rapidly generates a pool of mutant viruses, often resistant to the protease inhibitors [11,15].

HIV-1 protease, which is one of the most important targets of the highly topical research field known as computational virology [16], has been the subject of numerous molecular dynamics studies [17–25] and analysis of large crystallographic data sets [26–29]. These studies suggest that the most stable form of the enzyme is the semi-closed (or semi-open) one, followed by a more tightly closed form, whereas the enzyme in the open flap conformation is difficult to observe.

The aim of this work was to identify new ligands for the HIV-1 protease. We used a strategy based on a first phase of screening assisted by artificial intelligence (AI). The approach used in this phase was essentially the one described in [30], in which two neural networks work in concert. The first is a variational autoencoder (VAE), whose function is to generate a numerical representation of the molecules which can then be used by the second neural network. The latter is trained to associate the VAE numerical representation with a measure of the efficacy of known ligands of our target protein. Once trained, this system can be used for the rapid screening of large numbers of molecules. Subsequently, molecules identified as potential ligands were subjected to molecular docking, and the most interesting ones were validated by molecular dynamics. The advantage of this approach is that it is possible to eliminate many molecules that most likely have no chance of being interesting ligands of the target protein before proceeding with the molecular docking, thus reducing the computational load. Our most promising candidate belongs to a chemical class which, to the best of our knowledge, has not been considered for HIV-1 protease inhibition.

2. Materials and Methods

Molecules with known activity on the HIV-1 protease were obtained from ChEMBL [31–33] (query target CHEMBL243). The data set was analyzed with *pandas* [34,35] for the presence of duplicates, which were eliminated, and in this case, the one with the best inhibition value was kept.

The neural network (NN) used in this work consisted of two parts: a VAE and a deep feed-forward network (DNN), as described in [30]. NNs were implemented in Keras [36] with TensorFlow [37] as the backend. The VAE consists of two parts, namely the encoder and the decoder. The first one encodes a particular representation of the molecules (in our case, the SMILES representation [38]) in a numerical vector; the second one decodes the same vector in the starting representation. We used VAE because it has the advantage of minimizing the non-coding areas of the latent space (the latter is jargon to indicate the space containing vectors encoded by the VAE). Several VAE implementations are available [39–43], even already-trained ones, which can be used on SMILES strings; in this work, training was carried out as suggested in [42]. Vectors obtained by VAE staring from SMILES representations were used as input for various DNN architectures, which were trained to calculate the pChEMBL values associated with the corresponding molecules. We tried different DNNs, containing from three to nine fully connected layers (dense layers),

with or without dropout layers (up to six). The Adam algorithm was used for the optimizer, with mean absolute error as loss function, and the learning rate was controlled with the Keras ReduceLROnPlateau function by monitoring the loss function. Interested readers can find in the Supplementary Materials the pseudo-code to reproduce the DNN used to obtain the results described below. Once trained, the VAE-DNN system was used to predict the binding value (as pChEMBL score) on a data set containing 250,000 molecules (which we will refer to as ZINC250K), coded as non-isomeric SMILES, obtained from ZINC20 [44]; see also [43]. Molecules in ZINC250K with the best and worst predicted pChEMBL scores were selected; each of these two sets initially contained 933 items, and those that generated valid *pdbqt* files in Open Babel [45] were then used for further analysis.

Molecular docking was performed by means of AutoDock Vina 1.2.3 [46,47], essentially as described [48,49]. The PDB [50,51] structure 5IVQ [52] was used as receptor, and its *pdbqt* file was obtained by the AutoDockTools suite [53], with which the hydrogen atoms and Gasteiger-Marsili charges were added [54]. The docking box of dimensions (12.0 Å, 16.0 Å, 16.0 Å) was centered at the coordinates (18.6 Å, 18.6 Å, 6.6 Å).

Molecular dynamics has been performed in NAMD [55] in a water box with 15 Å padding, essentially as described [49,56,57] with few modifications. Briefly, a CHARMM36m force field [58] was used, and parameterization of the protein-ligand complex was carried out by means of CHARMM-GUI using Antechamber for ligand modeling [59–64]. Ionic strength and electroneutrality were obtained by adding potassium and chloride ions at a concentration of 150 mM. Periodic boundary conditions and the particle-mesh Ewald (PME) method have been used; the time step was 2 fs. Systems underwent 10,000 conjugate gradient minimization steps followed by 125,000 equilibration steps in canonical ensemble conditions, with the protein–ligand complex fixed, after which 10 ns production runs began in the NpT ensemble (Langevin dynamics at 303.15 K and Nosé-Hoover Langevin piston at 1.01325 bar). Structural analysis was conducted essentially as described in a VMD (version 1.9.3) environment [65–68].

Numerical calculations were performed using Numpy [69] and Scipy [70] in a Jupyter environment [71]. Graphs were obtained in Matplotlib [72]. The manipulation of SMILES strings, such as the conversion between isomeric and the non-isomeric forms, has been performed using the RDKit software suite [73].

3. Results

The main aim of this work was to try to identify new types of molecules capable of binding the HIV-1 protease. Virtual screening can be used in the early stages of the search for new protein ligands to speed up (even enormously) the evaluation of potential candidates. Currently, in the initial stages of in silico screening, molecular docking [74] is the first choice. Although docking is much faster than techniques that start from first principles such as molecular dynamics, being challenged with a large number of ligands requires extremely powerful computing infrastructures, and docking-dedicated databases today contain millions or even billions of entries [32,33,44,75–77].

In reality, most of the docking computing time is wasted by the algorithm to evaluate molecules of little chance to be ligands of the target of interest. Therefore, the computational strategy used in this work was based on the initial deployment of an NN for the rapid screening of potentially interesting molecules, consisting of a VAE and a DNN. At the VAE [30,78], previously trained with a sufficiently large database, molecules were presented as SMILES strings and then transformed by this into a numerical (i.e., vectorial) representation. In principle, any kind of molecular representation can be used for the autoencoder training, but representation as SMILES strings, despite being very simple, can obtain good results with a modest computational cost [39–43]. SMILES representation of molecules, whose experimental affinity for the HIV-1 protease is known, were obtained from ChEMBL. Only ligands with reported binding efficiency index (BEI), surface efficiency index (SEI), and pChEMBL value [79,80] were considered, and 5863 items in ChEMBL matched these criteria. After deleting duplicates and entries that cannot be processed

by the VAE (exclusion criterion was the length of the SMILES string of the ligand or the impossibility to obtain a latent vector),the final data set was composed of 4299 entries, each consisting of the VAE-generated vectorial representation of the molecule and the associated pChEMBL value. This data set was then split into training set and test set in a ratio of 0.8 to 0.2.

The first set was then used to train the DNN to transform the vector representation of molecules into pChEMBL values. Below, we will refer to what was obtained using an eight-layer DNN (see Section 2). This DNN is able to fit the train set almost perfectly: a linear relationship was obtained between the experimental and predicted pChEMBL values (slope 1.0000047, intercept—4.46^{-5}, R-value = 0.99999), which is not surprising, since the DNNs are capable of approximating any function [81], however complicated it may be. Obviously, the results on the test set are not so impressive, but a linear relationship between the experimental values and those estimated by the DNN is clearly visible, as reported in Figure 1. The linear fitting led to the following results: slope 0.58698, intercept 3.20243, R-value 0.64671, *p*-value 5^{-103}, standard error 0.02363, intercept standard error 0.18865.

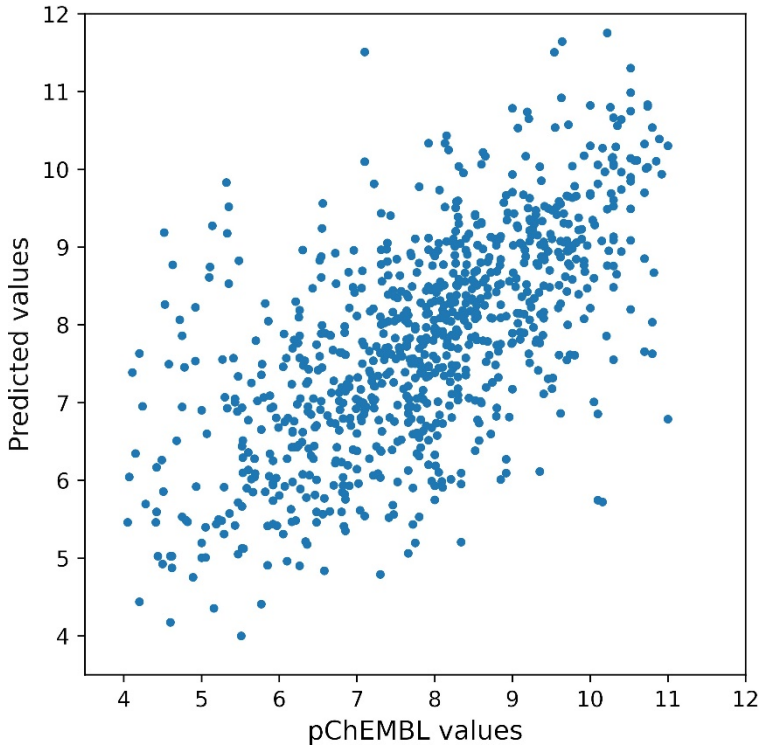

Figure 1. Results of fitting by the DNN described in the text on the ChEMBL test data set. The figure shows the prediction of the neural network described in the text on the ChEMBL test set (see also, Supplementay Data). Experimental data are reported on the horizontal axis, whereas the DNN predicted values are on the vertical one. The best fitting performed in Scipy gives as a result a straight line (not shown) with parameters: slope 0.58698, intercept 3.20243, R-value 0.64671, *p*-value 5^{-103}, standard error 0.02363, intercept standard error 0.18865.

No tendency to over-fitting (which could lead to a loss of ability to generalize by the DNN) was observed with any of the architectures used (not shown). The trained DNN was then used to predict the pChEMBL values of a larger data set. Here, we used the ZINC250K data set containing 250,000 molecules. The obtained pChEMBL scores range from a minimum of 3.02 to a maximum of 13.49. Interested readers can find the composition of this dataset in the Supplementay Data, where SMILES codes used for the calculation of the vector representation and the corresponding calculated score are reported.

The good performance of DNN on the test set, as reported above, does not necessarily imply that it is able to obtain similar results on random sets of molecules, such as the ZINC250K data set. This is because databases reporting experimental screening results for a particular target are most likely affected by bias. In general, an excess of particular classes of molecules, or functional groups, is expected in these data sets: if a class of molecule is known to be a good ligand for a particular target protein, many other molecules of the same class, or containing the same functional groups, will have been tested experimentally in an attempt to find new and better ligands, or in an attempt to find relationships between ligand structures and their activity (QSAR studies). However, even when there should not be this kind of bias, the chemical space is so enormous that a data set of a few thousand molecules is focused on only a small fraction of the possible ones (therefore, the data set is biased, inevitably; see Section 4).

Keeping what has been reported above in mind, we evaluated the validity our AI system predictions by considering how the predicted best (and worst) ligands performed in a completely different computational setup, namely molecular docking. We selected as best predicted ligands those with calculated pChEMBL values ≥ 10.0, and as worst predicted ligands, an equal number of molecules with the lowest predicted pChEMBL value (each of the two data sets containing 933 elements). These molecules were subjected to molecular docking on an HIV-1 protease template by means of AutoDock Vina, as detailed in Materials and Methods. The results of this analysis are shown in Figure 2, which reports the distribution of the calculated binding energies (BEs).

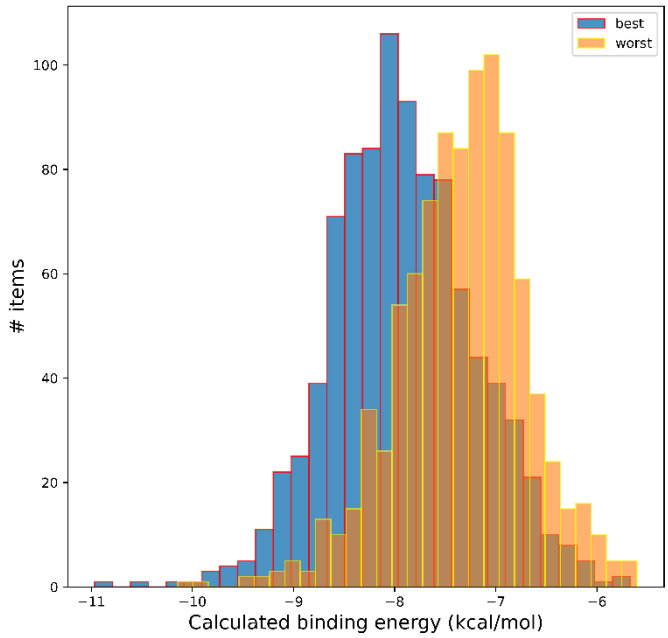

Figure 2. Molecular docking analysis of the best and worst predicted ligands by the DNN in the ZINC250K data set. The histogram shows the number of entries in the binned data of the two data sets, according to the calculated binding energies in the docking experiments. AutoDock Vina software suite was used for docking, as detailed in the main text. Numerical analysis was performed in Numpy.

Even if there is a certain overlap, the two distributions are clearly different, both as average and as extreme values. Molecules predicted as best ligands by the DNN are, on average, better than those predicted as bad ligands: the average BE of the best ligand subset is 7.92 kcal/mol (standard deviation 0.71), whereas for the worst subset, it is 7.33 kcal/mol (standard deviation 0.64). The Kolmogorov–Smirnov test, performed assuming as the null hypothesis that the two distributions are identical, returns as distance 0.38645 and p-value 2.15438^{-62}. This may be taken as evidence against the null hypothesis

and, consequently, that the two distributions are not identical. This suggests that the DNN we trained can be used effectively to select ligands which are more likely to be good ligands, avoiding wasting computational time on molecules that probably will not give any interesting outcome. This result is remarkable considering the diversity of the training data set obtained from ChEMBL and the ZINC250K data set.

To further validate the results obtained, we analyzed a series of molecules using molecular dynamics. We focused our attention on molecules that were part of the best set based on what was predicted by the DNN, and with a BE after molecular docking better than that of the ligand present in the 5IVQ structure. This ligand is a potent inhibitor, and probably a good candidate for clinical development; its chemical name is methyl N-[(2S)-1-[3-[(2R)-morpholin-2-yl]propylamino]-1-oxo-3,3-diphenylpropan-2-yl]carbamate [52]. Molecular docking of this molecule on its own receptor, carried out using the protocol reported in Section 2, leads to a calculated BE of 9.149 kcal/mol. Besides the BE value, we also manually evaluated the goodness of the docking obtained; it should be remembered that we used the procedure in which the receptor is rigid, so we carefully considered the presence of clashes and how extensively the ligand molecule occupied the region of the active site. By these criteria, we considered for molecular dynamics the following ZINC250K molecules: ZINC1040457718 (BE 10.965 kal/mol), ZINC948788229 (BE 10.534 kcal/mol), ZINC991374169 (BE 10.156 kcal/mol), ZINC987999904 (BE 9.642 kcal/mol). Moreover, we queried both ZINC20 and PubChem for molecules similar to ZINC1040457718 (the best ligand after molecular docking) using the built-in Tanimoto similarity search engines. We obtained 124 molecules which were then subjected to molecular docking on the HIV-1 protease (not shown). Many of these molecules have shown good affinity for the enzyme, but one showed a very interesting affinity (ZINC31942116, BE 12,231 kcal/mol), so we also considered this molecule for molecular dynamics. As a control, the dynamics of the holoenzyme and of the enzyme bound to its crystallographic ligand after docking were also performed.

After 10 ns of simulation, all the molecules considered were still in the active site of the enzyme. However, taking these data as significant evidence of the fact that we are dealing with a good ligand candidate is absolutely not sufficient. In our case, inspection of the molecular dynamics trajectories suggested a simple criterion to evaluate the effectiveness of the protein-ligand interaction: all the protein-ligand complexes considered, except one, showed a higher root mean squared deviation (RMSD) of the protein atoms than that obtained with the crystallographic ligand. Indeed, whereas in the case of the protease bound to its crystallographic ligand, an RMSD equal to 1.24 ± 0.46 Å was observed (the holoenzyme showed a similar RMSD to that of the crystallographic ligand, 1.23 ± 0.47 Å), in the case of the ligands listed above, the value was higher (for example, 1.40 ± 0.55 Å for ZINC987999904, 1.58 ± 0.69 Å for ZINC31942116), except for ZINC991374169, for which an RMSD of 1.28 ± 0.48 Å was obtained. However, even more interesting was the aspect of the protease, which remained in the closed conformation of the active site in the case of the crystallographic ligand and of ZINC991374169 (see Figure 3), whilst it assumed a swollen, semi-open conformation with all the other ligands (not shown), thus justifying the slightly higher value of the RMSD without a loss of overall stability of the protein. Furthermore, whereas the crystallographic ligand and ZINC991374169 after 10 ns were very close to the starting position (obtained by docking), as reported in Figure 3, in the case of the other molecules, the final position was very different, with a considerable mobility of these molecules into the semi-open active site during the simulation (not shown). All this suggests that only ZINC991374169 could be considered a plausible candidate for experimental validation. The structures of ZINC991374169 and of the crystallographic ligand are shown in Figure 3.

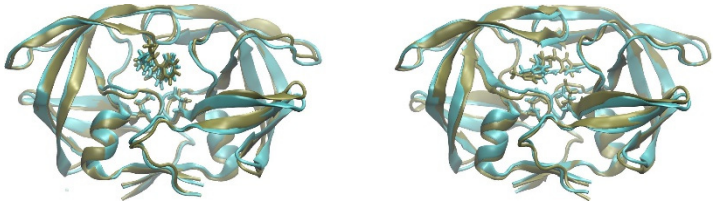

Figure 3. Structures of the protease–ligand complexes. The starting simulation structures of the protease–ligand complexes are reported in cyan, whereas those after 10 ns simulation are reported in tangerine. Starting simulation structures were obtained by minimization of the protease–ligand complex after molecular docking, as described in Section 2. The D25 residue of the HIV-1 protease is highlighted as licorice in all structures. Left structures refer to the HIV-1 protease bound to the crystallographic ligand reported in 5IVQ (see text); right structures refer to the HIV-1 protease bound to ZINC991374169.

The results of the first phase of screening were validated by means of classical techniques of computational biochemistry, i.e., docking and molecular dynamics (Figure 4).

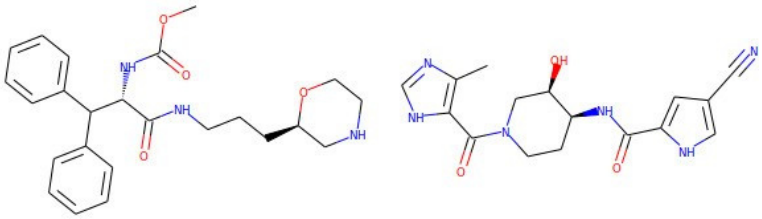

Figure 4. Structural formulas of molecules with best result after molecular dynamics. The structure of the crystallographic ligand present in the PDB entry 5IVQ is shown on the right; this molecule is reported in ZINC20 as ZINC584904731. On the left, the structure of the ligand identified in this work as best candidate, ZINC991374169.

4. Discussion

The search for new molecules capable of blocking the replication of pathogens, including viruses, has become a pressing clinical concern [8,9], and future projections indicate that if the current trend is not reversed, infectious diseases will once again be one of the main, if not the main, causes of death. Our interest has turned to one of the main drug targets for AIDS therapy, namely the HIV-1 protease. Several drugs are used today in clinical practice whose mechanism of action is based on the inhibition of this enzyme [11,13]. However, after years of use, even these molecules are beginning to be interested by resistance phenomena, with the serious possibility of a dramatic throwback to the early years of the AIDS pandemic. Very often, the search for new molecules capable of binding specific target proteins starts with computational methods that allow for the rapid screening of a large number of potential candidates. The vastness of the chemical space, or rather, the more than astronomical number of possible molecules, make the research always and only partial [82–84], hence the need for techniques capable of analyzing a large number of molecules quickly and with the least possible computing power. The new AI techniques that are increasingly gaining ground have the potential to greatly accelerate virtual screening processes [78].

We used a simple AI system for the initial screening of a data set containing a number of candidate molecules. Using this system, it has been possible to identify, in a short time, a new molecule with remarkable affinity in silico for the active site of the HIV-1 protease. As expected, this AI system can be used only for a first screening, but it should be noted, however, that this is a general problem: if we were really able to predict with absolute certainty whether a molecule is an effective drug, most of the problems of medicine would already be solved. However, it should be noted that computational analysis of protein-ligand interactions is an old and still not completely solved problem [85], as well as the evaluation of binding affinities and the effects on protein dynamics and functions. Bearing

in mind the above limitations, the results we obtained for ZINC991374169 are interesting, particularly when compared with that of the inhibitor bound to the HIV-1 protease used as a template for molecular docking (the PDB entry 5IVQ). Our results suggest that this molecule could be an interesting scaffold, which deserves to be explored.

Besides the particular molecule identified here, whose real efficacy as inhibitor of the HIV-1 protease will be established only experimentally, this work shows how it is possible to use AI-based computational techniques on protein targets to significantly reduce the search space in virtual screening. It is also interesting to underline the criterion (self-contained in a computational context) for the final acceptance of a good candidate ligand, i.e., the evaluation of the RMSD of the protein target compared to an adequate control. We wish to underline once again how this evaluation/validation criterion of potential ligands can be used when the structural (mainly crystallographic) data of control ligands whose activity is known are available. The availability of numerous viral protein structures, combined with the large number of experimental inhibition data, make the search for new antivirals an extremely interesting field of application for these computational techniques.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom13050858/s1, Figure S1: ChEMBL test data set and SMILES codes used for the calculation of the vector.

Author Contributions: Conceptualization, L.L.P.; methodology, L.L.P.; software, L.L.P.; validation, L.L.P., R.A., A.B., and L.S.; formal analysis, L.L.P.; investigation, L.L.P.; data curation, L.L.P.; writing—original draft preparation, L.L.P.; writing—review and editing, L.L.P., R.A., A.B., and L.S.; visualization, L.L.P.; project administration, L.L.P., R.A., and A.B.; funding acquisition, A.B., R.A., and L.S.; supervision and final approval A.B., and L.L.P.; critical revision of the manuscript for important intellectual content L.L.P. and A.B.; Finally, A.B. and R.A., equally contributed as co-first authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data obtained in this work and software not freely available online are reported in the text and in the Supplementary Materials. Any other data can be provided by the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NN	Neural network
DNN	Deep feed-forward network
VAE	Variational autoencoder
QSAR	Quantitative structure–activity relationship
RMSD	Root mean squared deviation

References

1. World Health Organization—HIV. Available online: https://www.who.int/news-room/fact-sheets/detail/hiv-aids (accessed on 17 December 2022).
2. Yamamoto, H.; Matano, T. Anti-HIV adaptive immunity: Determinants for viral persistence. *Rev. Med. Virol.* **2008**, *18*, 293–303. [CrossRef] [PubMed]
3. Ishii, H.; Matano, T. Development of an AIDS vaccine using Sendai virus vectors. *Vaccine* **2015**, *33*, 6061–6065. [CrossRef] [PubMed]
4. Burton, D.R. Advancing an HIV vaccine; advancing vaccinology. *Nat. Rev. Immunol.* **2019**, *19*, 77–78. [CrossRef] [PubMed]
5. Volberding, P.A.; Deeks, S.G. Antiretroviral therapy and management of HIV infection. *Lancet* **2010**, *376*, 49–62. [CrossRef]
6. Tozser, J. Stages of HIV replication and targets for therapeutic intervention. *Curr. Top. Med. Chem.* **2003**, *3*, 1447–1457. [CrossRef]

7. Rhee, S.Y.; Gonzales, M.J.; Kantor, R.; Betts, B.J.; Ravela, J.; Shafer, R.W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298–303. [CrossRef]
8. Clutter, D.S.; Jordan, M.R.; Bertagnolio, S.; Shafer, R.W. HIV-1 drug resistance and resistance testing. *Infect. Genet. Evol.* **2016**, *46*, 292–307. [CrossRef]
9. Hamers, R.L.; Rinke de Wit, T.F.; Holmes, C.B. HIV drug resistance in low-income and middle-income countries. *Lancet HIV* **2018**, *5*, e588–e596. [CrossRef]
10. Parbie, P.K.; Abana, C.Z.Y.; Kushitor, D.; Asigbee, T.W.; Ntim, N.A.A.; Addo-Tetebo, G.; Ansong, M.R.D.; Ofori, S.B.; Mizutani, T.; Runtuwene, L.R.; et al. High-level resistance to non-nucleos(t)ide reverse transcriptase inhibitor based first-line antiretroviral therapy in Ghana; A 2017 study. *Front. Microbiol.* **2022**, *13*, 2953. [CrossRef]
11. Weber, I.T.; Wang, Y.F.; Harrison, R.W. HIV protease: Historical perspective and current research. *Viruses* **2021**, *13*, 839. [CrossRef]
12. Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249–284. [CrossRef]
13. Konvalinka, J.; Kräusslich, H.G.; Müller, B. Retroviral proteases and their roles in virion maturation. *Virology* **2015**, *479*, 403–417. [CrossRef]
14. Mótyán, J.A.; Miczi, M.; To˝zsér, J. Dimer interface organization is a main determinant of intermonomeric interactions and correlates with evolutionary relationships of retroviral and retroviral-like Ddi1 and Ddi2 proteases. *Int. J. Mol. Sci.* **2020**, *21*, 1352. [CrossRef] [PubMed]
15. Matthew, A.N.; Leidner, F.; Lockbaum, G.J.; Henes, M.; Zephyr, J.; Hou, S.; Rao, D.N.; Timm, J.; Rusere, L.N.; Ragland, D.A.; et al. Drug design strategies to avoid resistance in direct-acting antivirals and beyond. *Chem. Rev.* **2021**, *121*, 3238–3270. [CrossRef] [PubMed]
16. Machado, M.R.; Pantano, S. Fighting viruses with computers, right now. *Curr. Opin. Virol.* **2021**, *48*, 91–99. [CrossRef] [PubMed]
17. Chatfield, D.C.; Brooks, B.R. HIV-1 protease cleavage mechanism elucidated with molecular dynamics simulation. *J. Am. Chem. Soc.* **1995**, *117*, 5561–5572. [CrossRef]
18. Liu, H.; Müller-Plathe, F.; van Gunsteren, W.F. A combined quantum/classical molecular dynamics study of the catalytic mechanism of HIV protease. *J. Mol. Biol.* **1996**, *261*, 454–469. [CrossRef]
19. Wang, W.; Kollman, P.A. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *J. Mol. Biol.* **2000**, *303*, 567–582. [CrossRef]
20. Hornak, V.; Okur, A.; Rizzo, R.C.; Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 915–920. [CrossRef]
21. Carnevale, V.; Raugei, S.; Neri, M.; Pantano, S.; Micheletti, C.; Carloni, P. Multi-scale modeling of HIV-1 proteins. *J. Mol. Struct. THEOCHEM* **2009**, *898*, 97–105. [CrossRef]
22. Sadiq, S.K.; Muniz Chicharro, A.; Friedrich, P.; Wade, R.C. Multiscale approach for computing gated ligand binding from molecular dynamics and Brownian dynamics simulations. *J. Chem. Theory Comput.* **2021**, *17*, 7912–7929. [CrossRef] [PubMed]
23. Lockbaum, G.J.; Leidner, F.; Rusere, L.N.; Henes, M.; Kosovrasti, K.; Nachum, G.S.; Nalivaika, E.A.; Ali, A.; Kurt Yilmaz, N.; Schiffer, C.A. Structural adaptation of darunavir analogues against primary mutations in HIV-1 protease. *ACS Infect. Dis.* **2019**, *5*, 316–325. [CrossRef] [PubMed]
24. Tang, W.S.; da Silva, G.M.; Kirveslahti, H.; Skeens, E.; Feng, B.; Sudijono, T.; Yang, K.K.; Mukherjee, S.; Rubenstein, B.; Crawford, L. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS Comput. Biol.* **2022**, *18*, e1010045. [CrossRef] [PubMed]
25. Kaynak, B.T.; Krieger, J.M.; Dudas, B.; Dahmani, Z.L.; Costa, M.G.; Balog, E.; Scott, A.L.; Doruker, P.; Perahia, D.; Bahar, I. Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods. *Front. Mol. Biosci.* **2022**, *9*, 832847. [CrossRef]
26. Teodoro, M.L.; Phillips, G.N., Jr.; Kavraki, L.E. Understanding protein flexibility through dimensionality reduction. *J. Comput. Biol.* **2003**, *10*, 617–634. [CrossRef]
27. Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R.L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321–330. [CrossRef]
28. Palese, L.L. Conformations of the HIV-1 protease: A crystal structure data set analysis. *Biochim. Biophys. Acta Proteins Proteom.* **2017**, *1865*, 1416–1422. [CrossRef]
29. Palese, L.L. Analysis of the conformations of the HIV-1 protease from a large crystallographic data set. *Data Brief* **2017**, *15*, 696–700. [CrossRef]
30. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef]
31. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
32. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J.P. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620. [CrossRef] [PubMed]

33. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940. [CrossRef] [PubMed]
34. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
35. Pandas Development Team. Pandas-dev/Pandas: Pandas. Available online: https://zenodo.org/record/7857418#.ZGM8ZnZByUk (accessed on 13 April 2023).
36. Chollet, F.; Pal, S. Keras. Available online: https://keras.io (accessed on 21 December 2022).
37. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available online: https://www.tensorflow.org/ (accessed on 13 April 2023).
38. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
39. Bjerrum, E.J.; Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **2018**, *8*, 131. [CrossRef]
40. Duvenaud, D.; Hirzel, T. Molecular Autoencoder. Available online: https://github.com/HIPS/molecule-autoencoder (accessed on 21 December 2022).
41. Hodak, M.; Pechersky, Y.; Yi, H.; Rahman, F. A Keras Implementation of Aspuru-Guzik Molecular Autoencoder Paper. Available online: https://github.com/maxhodak/keras-molecules (accessed on 21 December 2022).
42. Félix, E. Autoencoder Ipython. Available online: https://github.com/chembl/autoencoder_ipython (accessed on 21 December 2022).
43. Sanchez-Lengeling, B. Chemical VAE. Available online: https://github.com/aspuru-guzik-group/chemical_vae (accessed on 21 December 2022).
44. Irwin, J.J.; Tang, K.G.; Young, J.; Dandarchuluun, C.; Wong, B.R.; Khurelbaatar, M.; Moroz, Y.S.; Mayfield, J.; Sayle, R.A. ZINC20—A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073. [CrossRef] [PubMed]
45. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 1–14. [CrossRef] [PubMed]
46. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
47. Eberhardt, J.; Santos-Martins, D.; Tillack, A.F.; Forli, S. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891–3898. [CrossRef]
48. Isgrò, C.; Sardanelli, A.M.; Palese, L.L. Systematic search for SARS-CoV-2 main protease inhibitors for drug repurposing: Ethacrynic acid as a potential drug. *Viruses* **2021**, *13*, 106. [CrossRef]
49. Sardanelli, A.M.; Isgrò, C.; Palese, L.L. SARS-CoV-2 main protease active site ligands in the human metabolome. *Molecules* **2021**, *26*, 1409. [CrossRef]
50. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
51. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G.V.; Christie, C.H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J.M.; et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021**, *49*, D437–D451. [CrossRef] [PubMed]
52. Bungard, C.J.; Williams, P.D.; Ballard, J.E.; Bennett, D.J.; Beaulieu, C.; Bahnck-Teets, C.; Carroll, S.S.; Chang, R.K.; Dubost, D.C.; Fay, J.F.; et al. Discovery of MK-8718, an HIV protease inhibitor containing a novel morpholine aspartate binding group. *ACS Med. Chem. Lett.* **2016**, *7*, 702–707. [CrossRef] [PubMed]
53. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef] [PubMed]
54. Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19*, 3181–3184. [CrossRef]
55. Phillips, J.C.; Hardy, D.J.; Maia, J.D.; Stone, J.E.; Ribeiro, J.V.; Bernardi, R.C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130. [CrossRef]
56. Bossis, F.; Palese, L.L. Amyloid beta (1–42) in aqueous environments: Effects of ionic strength and E22Q (Dutch) mutation. *Biochim. Biophys. Acta Proteins Proteom.* **2013**, *1834*, 2486–2493. [CrossRef]
57. Bossis, F.; De Grassi, A.; Palese, L.L.; Pierri, C.L. Prediction of high-and low-affinity quinol-analogue-binding sites in the aa 3 and bo 3 terminal oxidases from Bacillus subtilis and Escherichia coli. *Biochem. J.* **2014**, *461*, 305–314. [CrossRef]
58. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B.L.; Grubmüller, H.; MacKerell, A.D., Jr. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73. [CrossRef]
59. Jo, S.; Kim, T.; Iyer, V.G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865. [CrossRef]
60. Brooks, B.R.; Brooks, C.L., III; Mackerell, A.D., Jr.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614. [CrossRef]

61. Lee, J.; Cheng, X.; Swails, J.M.; Yeom, M.S.; Eastman, P.K.; Lemkul, J.A.; Wei, S.; Buckner, J.; Jeong, J.C.; Qi, Y.; et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* **2016**, *12*, 405–413. [CrossRef] [PubMed]
62. Kim, S.; Lee, J.; Jo, S.; Brooks, C.L., III; Lee, H.S.; Im, W. CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.* **2017**, *38*, 1879–1886. [CrossRef] [PubMed]
63. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef]
64. Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260. [CrossRef] [PubMed]
65. Bossis, F.; Palese, L.L. Molecular dynamics in cytochrome c oxidase Mössbauer spectra deconvolution. *Biochem. Biophys. Res. Commun.* **2011**, *404*, 438–442. [CrossRef]
66. Palese, L.L. Correlation analysis of Trp-cage dynamics in folded and unfolded states. *J. Phys. Chem. B* **2015**, *119*, 15568–15573. [CrossRef]
67. Palese, L.L. Random matrix theory in molecular dynamics analysis. *Biophys. Chem.* **2015**, *196*, 1–9. [CrossRef]
68. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]
69. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
70. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
71. Jupyter. Available online: https://jupyter.org/ (accessed on 21 December 2022).
72. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
73. RDKit: Open-Source Cheminformatics. Available online: https://www.rdkit.org (accessed on 19 December 2022).
74. Pagadala, N.S.; Syed, K.; Tuszynski, J. Software for molecular docking: A review. *Biophys. Rev.* **2017**, *9*, 91–102. [CrossRef] [PubMed]
75. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef] [PubMed]
76. Irwin, J.J.; Shoichet, B.K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182. [CrossRef] [PubMed]
77. Sterling, T.; Irwin, J.J. ZINC15—Ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef]
78. Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **2019**, *119*, 10520–10594. [CrossRef]
79. Abad-Zapatero, C.; Metz, J.T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* **2005**, *10*, 464–469. [CrossRef]
80. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [CrossRef]
81. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
82. Polishchuk, P.G.; Madzhidov, T.I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679. [CrossRef] [PubMed]
83. Reymond, J.L. The chemical space project. *Acc. Chem. Res.* **2015**, *48*, 722–730. [CrossRef] [PubMed]
84. Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861. [CrossRef] [PubMed]
85. Gilson, M.K.; Zhou, H.X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42. [CrossRef] [PubMed]