**Supporting Information**

**Enhancing HIV-1 Protease Inhibitor Design using Interpretable Machine Learning and Molecular Modeling**

Hathaichanok Chuntakaruka,e, Kajjana Boonpalitb, Jiramet Kinchagawatb, Fahsai Nakarinb,

Kowit Hengphasatpornc\*, Yasuteru Shigetac, Tanatorn Khotavivattanad, Chanat Aonbangkhendd,

Sarana Nutanongb\*, Thanyada Rungrotmongkola,e\* and Supot Hannongbuaa,f\*

a Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University,

Bangkok 10330, Thailand

b School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology

(VISTEC), Rayong 21210, Thailand

c Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577,

Japan

d Center of Excellence in Natural Products Chemistry, Department of Chemistry, Faculty of Science,

Chulalongkorn University, Bangkok 10330, Thailand

e Center of Excellence in Structural and Computational Biology, Department of Biochemistry, Faculty of

Science, Chulalongkorn University, Bangkok 10330, Thailand

f Center of Excellence in Computational Chemistry (CECC), Department of Chemistry, Faculty of Science,

Chulalongkorn University, Bangkok 10330, Thailand

E-mail: supot.h@chula.ac.th; [thanyada.r@chula.ac.th](mailto:thanyada.r@chula.ac.th); snutanon@vistec.ac.th; kowith@ccs.tsukuba.ac.jp

Phone: +66-2218-7707-6

***Feature importance analysis***

The Gini importance, also referred to as the mean decrease of the Gini index (MDG), quantifies the influence of a feature on the model's prediction. The feature with the highest Gini importance is considered the most important as it has the largest impact on the model's performance. The Gini importance of each feature is determined by calculating the average of the total weighted impurity reduction caused by the splitting at the node where that feature is used.

where *p*(*t*) is proportion of samples reaching node *t* compared to total samples, *v*(*st*) is a feature used in splitting *st*, and ∆*i*(*st*,*t*) is the impurity reduction caused by splitting *st* at node *t*

Additionally, the analysis of the top-ranked essential descriptors was performed using RDKit 30. These descriptors 19 were then converted into chemical substructures. The substructures represented by these descriptors were matched with the structures of the inhibitors. Consequently, the substructures could provide clarification on whether they consisted of active or inactive inhibitors, thus guiding rational drug design.

***Dimensionality reduction techniques***

Dimensionality reduction was performed on the original molecular fingerprint using principal component analysis (PCA) and T-distributed stochastic neighbor embedding (T-SNE). Theoretically, PCA aims to preserve the global structure of the dataset by preserving large pairwise distances to maximize variance, while T-SNE focuses on preserving the local structure and similarities between neighboring molecules based on the Jaccard distance. This process reduced the high-dimensional features to two dimensions: PCA-0/PCA-1 and T-SNE0/T-SNE1. The purpose was to visualize the chemical space of the dataset. For PCA, we used the default hyperparameters implemented in the scikit-learn library. As for T-SNE, we set the perplexity to ten and utilized the Jaccard distance as the metric for T-SNE.

A picture containing text, printing, screenshot, typography

Description automatically generated**Figure S1.** Design 474 DRV analogs by modifying functional groups at P1, P1', P2, and P2' on DRV. The heatmap shows the change in binding free energy(= − ) compared to of DRV.

A colorful molecule structure on a white background

Description automatically generated with low confidence

**Figure S2.** Superimposition of the template of HIV-1 PR/DRV (purple) complex and the docked structure (green) obtained from the GNINA 1.0 molecular docking program visualized by UCSF Chimera package

(A)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

(B)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

(C)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

(D)

**A picture containing text, colorfulness, screenshot, diagram

Description automatically generated**

(E)

**A picture containing screenshot, colorfulness, text, diagram

Description automatically generated**

(F)

**A picture containing text, screenshot, diagram, colorfulness

Description automatically generated**

(G)

**A picture containing text, colorfulness, screenshot, diagram

Description automatically generated**

(H)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

(I)

**A picture containing colorfulness, screenshot, diagram, green

Description automatically generated**

(J)

**A picture containing text, colorfulness, screenshot, diagram

Description automatically generated**

(K)

**A picture containing screenshot, colorfulness, text, diagram

Description automatically generated**

(L)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

(M)

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

**Figure S3** Interactions between 90 candidates and various forms of HIV-1 protease (PR) were assessed, including the wild-type (A), as well as specific variants such as D30N (B), V32I (C), M46L (D), G48V (E), I50V (F), I54M (G), I54V (H), L76V (I), V82A (J), I84V (K), N88S (L), and L90M (M).