

SHORT COMMUNICATION

©Ivanova, Skvortsov

THE PREDICTION OF MAIN PROTEASE SARS-CoV-2 INHIBITION BASED ON MODELS OF ENZYME-INHIBITOR COMPLEXES

Ya.O. Ivanova*, V.S. Skvortsov

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: yana.emris@gmail.com

A set of linear regression equations predicting the IC_{50} values for SARS-CoV-2 main protease inhibitors was analyzed. For 180 competitive inhibitors, we have simulated the molecular dynamics of enzyme-inhibitor complexes with known structures or modeled using molecular docking. In the docking procedure, the selection of final poses was restricted by similarity to known structural analogs. The values of the energy contributions obtained by means of calculation of the free energy change of the enzyme-inhibitor complex performed by two variants of the MMPBSA (MMGBSA) method and a number of physicochemical characteristics of the inhibitors were used as independent variables. During the learning process, indicator variables were used for inhibitor subsets obtained from various literature sources to compensate the existing systematic deviations from the target value. A leave one out and leave 20% out cross validation procedures were used to evaluate the prediction quality. For the total logarithmic range width of 3.71, the mean error in predicting the $\lg(IC_{50})$ value was 0.45 log units. The stability of the prediction depending on the variability of the complex in molecular dynamics was investigated.

Key words: SARS-CoV-2; main protease; competitive inhibitors; QSAR

DOI: 10.18097/PBMC20236905322

INTRODUCTION

The present study is a continuation of our previous work [1], in which we showed the importance of restricting the ligand position at the binding site during the docking procedure. In our previous work, unsupervised docking was initially used, where the selection of the “best” position was based only on the evaluation function, and the ligand position was controlled *ex post facto*. Currently, several docking methods exist that control ligand positioning order to find a position closer to the known structural analog. This avoided missing data when the solution for a group of closely related molecules did not match the data obtained from the analysis of crystallographic structures or the general trend for most members of the group. In the previous work, there were 70 such solutions out of 146. Also, the present study included an additional dataset [2] on competitive inhibitors of SARS-CoV-2 main protease (SARS-CoV-2 M^{pro}) with known IC_{50} values, which was added to the general set. We have made certain modifications to the data preparation and added a new calculation procedure using the Amber 19 package [3]. As before, the main objective of this work was to analyze simple linear equations based on the use of data on known or modeled complexes of SARS-CoV-2 M^{pro} with its inhibitors.

MATERIALS AND METHODS

The dataset, including 146 competitive inhibitors of SARS-CoV-2 M^{pro} (dataset S1) selected in our previous work [1] has been used in this study.

The structures of the enzyme-inhibitor complex currently known for 34 compounds of this dataset are available in the Protein Data Bank [4]. Thirty-five compounds from [2] were added to this dataset. For one of them (L014, the compound ID is given in Supplementary Materials), previously included in sample S1, crystal structure data of the enzyme-inhibitor complex are now available (PDB ID: 7LMF). The IC_{50} data for all added compounds were obtained using the same substrate HiyteFluor-488ESATLQSGLRKAK-(QXL)- NH_2 (dataset S2). The range of $\lg(IC_{50})$ values for the entire set was 3.71 logarithmic units (l.u.; 1.26 to 4.97; Fig. 1). For the S2 set, the range of values was narrower (2.69 l.u.). A complete set of data on structures, distribution by structural group, the IC_{50} values, literature references, etc. are available in the Supplementary Materials to this article.

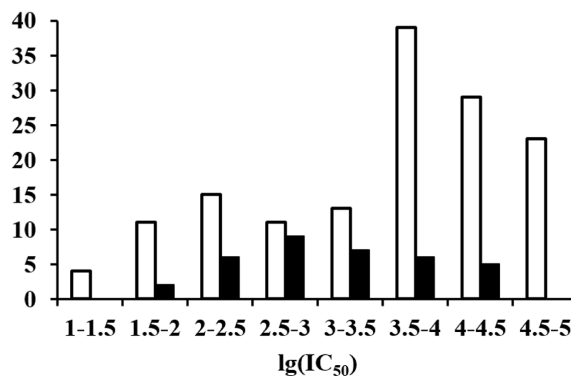


Figure 1. Distribution of $\lg(IC_{50})$ values for a set of compounds with known inhibitory activity against SARS-CoV-2 M^{pro} . White bars represent the S1 dataset, black bars represent the S2 dataset.

The initial modeling of enzyme-inhibitor complexes was performed as in [1] using the Schrodinger program [5]. However, in contrast to the previous study, for each of the previously obtained structural groups, an inhibitor molecule with a known crystal structure of its complex with SARS-CoV-2 M^{pro} was defined as a prototype. There was a restriction on the selection of structures: only those structures that had a similar position in the docking process were selected. The 3D structure of the complex of one of the group inhibitors taken from the PDB was used as the protein structure (binding site) for docking for each group. If no inhibitor with a known crystal structure was available for a given group, the matching inhibitor binding pose variants maximally represented in the group were used as the priority structure of the enzyme-inhibitor complex, and the SARS-CoV-2 M^{pro} structure variant from the complex with compound L039 (PDB ID: 7N44) was used as the binding site. All of the SARS-CoV-2 M^{pro} structures that were used were pre-aligned in space with respect to each other. The OPLS3e force field [6] was used to optimize the overall structure of the complexes. This procedure has been performed for all complexes. This included structures derived from the PDB. Free energy changes for each enzyme-inhibitor complex were calculated using the MMGBSA method (VSGB solvation model) as in [1]. A set of 7 energy terms was used as independent variables (group E1) to generate linear regression equations. These included the change in Coulomb interaction energy, energy of covalent interactions in the ligand and receptor, energy of van der Waals interactions, non-polar contribution to solvation energy by surface area, electrostatic contribution to solvation energy based on the generalized Born model, hydrogen bonding contribution, and contribution related to lipophilicity. In contrast to our previous work [1], as a next step of the complex structure optimization we additionally used a set of sequential molecular dynamics simulations (via the scheme described earlier [2]) using the Amber19 software package [3]. The last step

of the molecular dynamics simulation served as the basis for calculating the energy contributions to the free energy change of the complexes using the MMPBSA method [8]. A set of these parameters (E2) has also been used as an independent variable. This set included: (i) changes in the value of electrostatic and van der Waals interactions; (ii) hydrophobic and solvation contributions to the free energy change calculated by the Poisson-Boltzmann method and (iii) similar contributions calculated by the Generalized Born method; (iv) translational, rotational and vibrational entropic contributions [8].

The $\lg(\text{IC}_{50})$ value was used as the target for the prediction equations. The IC_{50} value was taken in nM. In addition to the energy parameters characterizing the properties of the complex, a set of 6 parameters (P) characterizing the properties of the ligand itself (molecular weight, total and polar volume, total and polar surface area, number of bonds on which rotation is possible) was calculated for each of the ligands by means of the Sybyl-X software [9]. These parameters were also used as independent variables. In the process of selecting linear regression equations, we varied both the set of independent variables and the composition of the set of observations. As in [1], the quality of the models was evaluated by the results of the leave-one-out cross-validation procedure; in addition, we used the leave-20%-out test. For this purpose, the data were sorted by IC_{50} value and then every *n*th element was selected, forming 5 samples with similar ranges of IC_{50} values.

RESULTS AND DISCUSSION

The variations in the position of the inhibitors at the SARS-CoV-2 M^{pro} binding site are quite large even when the structures of the complexes from the PDB have been aligned with each other (Fig. 2A). At the same time, using the chosen docking procedure it is possible to obtain a similar position of the inhibitors within their structural group (Fig. 2B,C).

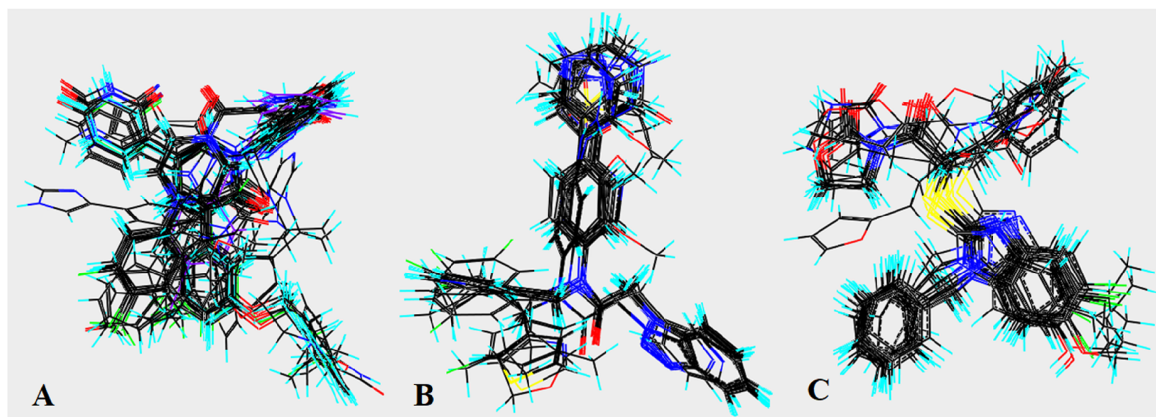


Figure 2. Position of the inhibitors within the known structure of their complexes with SARS-CoV-2 M^{pro} after alignment of the protein structures from the PDB (A); and alignment for inhibitor datasets from [2] (B) and [12] (C), obtained by docking to the SARS-CoV-2 M^{pro} structure. The orientation of alignments A, B, and C do not coincide.

The statistical characteristics of all models (equations) for predicting the $\lg(IC_{50})$ value discussed below are shown in the Table. In all cases, the possibility of eliminating observation outliers was not considered, although significant deviations (up to 2.22 l.u.) were observed for some of the PDB structures in the learning procedure. As a first option, we varied all 3 groups of variables (P, E1 and E2) for the dataset (S1, 145 observations) used in [1]. The data for compound L014, which was also included in the group of added inhibitors (dataset S2, 35 observations), were removed from dataset S1. None of the three groups of variables alone was sufficient to construct equations with acceptable performance (see equations #1, #2, and #3 in the Table). However, the procedure of randomly mixing the target value shows that even in this variant the equations have weak predictive power. It cannot be considered as significant because the significance criterion in the leave-one-out procedure is usually considered significant when the Q^2 value is greater than 0.6. The equations constructed for the combined sets of variables (#4, #5, and #6) had a significant predictive power. At the same time, when we used the S2 dataset as a test sample, the results did not meet the significance criteria. However, the dataset S2 has a narrower range of target values, and the linear regression equations for the same set of independent variables (equations #7, #8, and #9 in the Table) had a poor performance, even if we ignored the fact that they were irrelevant with this number of variables and these parameters (see features under random mixing of the target function). In addition, the predicted $\lg(IC_{50})$ values for the S2 dataset have a pronounced systematic error (Fig. 3A). The IC_{50} values obtained in different laboratories using the same set of compounds may differ even when using the same experimental technique. The best case is when different datasets have a certain number of the same compounds, and then the data can be equalized [10]. In this case, there are no such overlaps for our sample collected from literature sources. One workaround is to use indicator variables in learning that flag groups of data from a single source. This essentially shifts the target value by a fixed amount. In general, the fewer such indicator variables there are, the less likely overtraining will occur. We analyzed the dataset S1, and identified 2 subsets collected from the works [11, 12], for which the introduction of an indicator variable was relevant (example in Fig. 3B). These two indicator variables are labeled in the Table as the parameter set I1 (I2 is the indicator variable for dataset S2). When I1 was used in combination with data sets P, E1 and/or E2, the performance of the equations improved (#10, #11, and #12). However, because the range of target values in the test sample S2 was narrower than in the training dataset S1, the scatter of the predicted values was also quite large. Since the combined datasets S1 and S2 also gave slightly worse results using the indicator variable I2

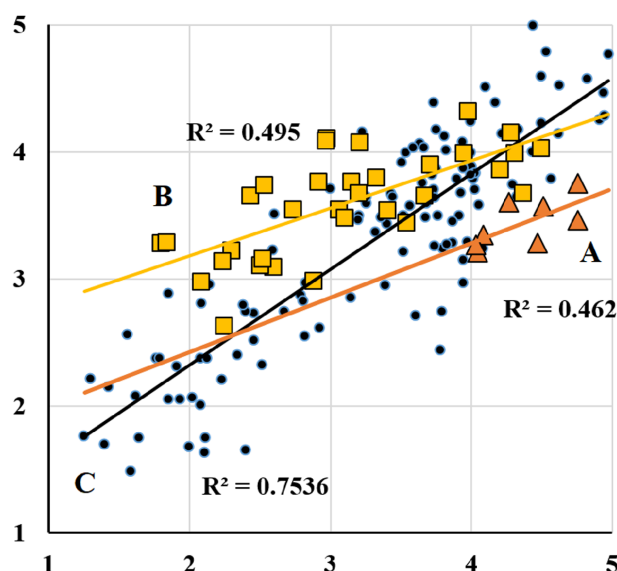


Figure 3. The comparison of experimentally determined and predicted values of $\lg(IC_{50})$ for the datasets from [11] (A, triangle), [2] (B, square), and the learning dataset (C). The prediction was performed using the modified equation 6 (Table), recalculated after removing the observations from [11] and [12] from the dataset. The abscissas axis shows the experimentally determined values of $\lg(IC_{50})$. The vertical axis shows the predicted values.

(#13, #14, and #15), the reason could be most likely attributed to the “noisiness” of the S2 dataset. Nevertheless, the overall parameters of the equations met the significance criterion of predictive power. As an additional test of quality, we present data for the leave-20%-out cross-validation procedure (#14.1–#14.5). The average R^2 for the test was 0.65 and the average error was 0.45 l.u. (total range of values 3.71 l.u.).

The set of independent variables used for equation 14 included variables from sets P and E2. When selecting the best of the total variables, the equations always included the set I1, the polar surface area and the polar volume from the set P, and from the set E2, the values of the electrostatic and van der Waals interactions and the solvation contribution calculated by the Poisson-Boltzmann method. Depending on the subset, the equations sometimes used I2, the total volume or molecular weight from P, the hydrophobic contribution calculated by Generalized Born, and the translational, rotational, and vibrational entropic contributions from E2.

Since some of the variables are correlated with each other, they may substitute for each other in different variants. The variable I2 was used in the general equation 14 and in two of the five tests (#14.1 and #14.5). It can be assumed that the S2 dataset is heterogeneous, but no other evidence was found.

The sets of variables E1 and E2 essentially characterized the same parameters of the complexes, but when using the combination of E1 and E2, the final

Table. The parameters of the $\lg(\text{IC}_{50})$ prediction equations obtained during learning and the results of testing

#	Number of observations (learning)	Variable set	Number of variables + constant, best equation (total)	R^2_L	ME_L	MaxE_L	R^2_{rand}	ME_{rand}	Q^2_{loo}	ME_{loo}	MaxE_{loo}	Number of observations (testing)	R^2_{test}	ME_{test}	$\text{MaxE}_{\text{test}}$
1	145	P	6(7)	0.49	0.57	2.22	0.05	0.78	0.44	0.60	2.30	35	0.01	0.73	1.86
2	145	E1	4(7)	0.51	0.56	1.94	0.02	0.78	0.48	0.57	2.04	35	0.24	1.42	3.73
3	145	E2	8(10)	0.49	0.57	1.59	0.06	0.77	0.43	0.61	1.85	35	0.07	0.62	1.67
4	145	P+E1	7(13)	0.71	0.41	1.45	0.05	0.77	0.67	0.43	1.57	35	0.32	1.80	3.70
5	145	P+E2	8(16)	0.73	0.39	1.46	0.07	0.78	0.71	0.41	1.50	35	0.37	0.76	1.99
6	145	P+E1+E2	11(22)	0.77	0.37	1.39	0.08	0.77	0.73	0.40	1.48	35	0.41	1.20	2.49
7	35	P+E1	7(13)	0.62	0.36	0.87	0.18	0.53	0.46	0.45	1.07	—	—	—	—
8	35	P+E2	5(16)	0.61	0.36	1.16	0.19	0.53	0.51	0.42	1.26	—	—	—	—
9	35	P+E1+E2	11(22)	0.75	0.28	0.93	0.42	0.45	0.56	0.40	1.09	—	—	—	—
10	145	P+E1+I1	6(15)	0.76	0.38	1.47	0.04	0.79	0.74	0.39	1.49	35	0.27	0.75	2.17
11	145	P+E2+I1	8(18)	0.76	0.38	1.23	0.09	0.76	0.74	0.40	1.25	35	0.19	0.55	1.58
12	145	P+E1+E2+I1	9(24)	0.80	0.36	1.18	0.05	0.78	0.77	0.38	1.25	35	0.35	0.59	1.73
13	180	P+E1+I1+I2	7(16)	0.67	0.44	1.70	0.06	0.76	0.64	0.46	1.75	—	—	—	—
14	180	P+E2+I1+I2	7(19)	0.71	0.41	1.37	0.07	0.77	0.69	0.42	1.40	—	—	—	—
15	180	P+E1+E2+I1+I2	8(25)	0.73	0.41	1.22	0.80	0.76	0.70	0.43	1.33	—	—	—	—
14.1	144	P+E2+I1+I2	11(19)	0.73	0.39	1.48	0.07	0.76	0.69	0.42	1.56	36	0.70	0.42	1.22
14.2	144	P+E2+I1+I2	8(19)	0.73	0.40	1.12	0.05	0.77	0.70	0.42	1.16	36	0.64	0.45	1.47
14.3	144	P+E2+I1+I2	8(19)	0.72	0.39	1.45	0.05	0.78	0.69	0.41	1.49	36	0.69	0.46	1.23
14.4	144	P+E2+I1+I2	8(19)	0.71	0.41	1.31	0.04	0.78	0.68	0.43	1.37	36	0.68	0.46	1.28
14.5	144	P+E2+I1+I2	9(19)	0.73	0.40	1.18	0.09	0.76	0.70	0.42	1.22	36	0.57	0.46	1.56

R^2_L – R^2 of learning; ME_L – Mean error of learning; MaxE_L – Maximum error of learning; R^2_{rand} – Mean R^2 in the learning procedure by random-mixed-value set; ME_{rand} – Mean error in the learning procedure by random-mixed-value set; Q^2_{loo} – Q^2 of the model in the leave-one-out procedure; ME_{loo} – Mean error in the leave-one-out procedure; MaxE_{loo} – Maximum error in the leave-one-out procedure; R^2_{test} – R^2 of testing; ME_{test} – Mean error of testing; $\text{MaxE}_{\text{test}}$ – Maximum error of testing.

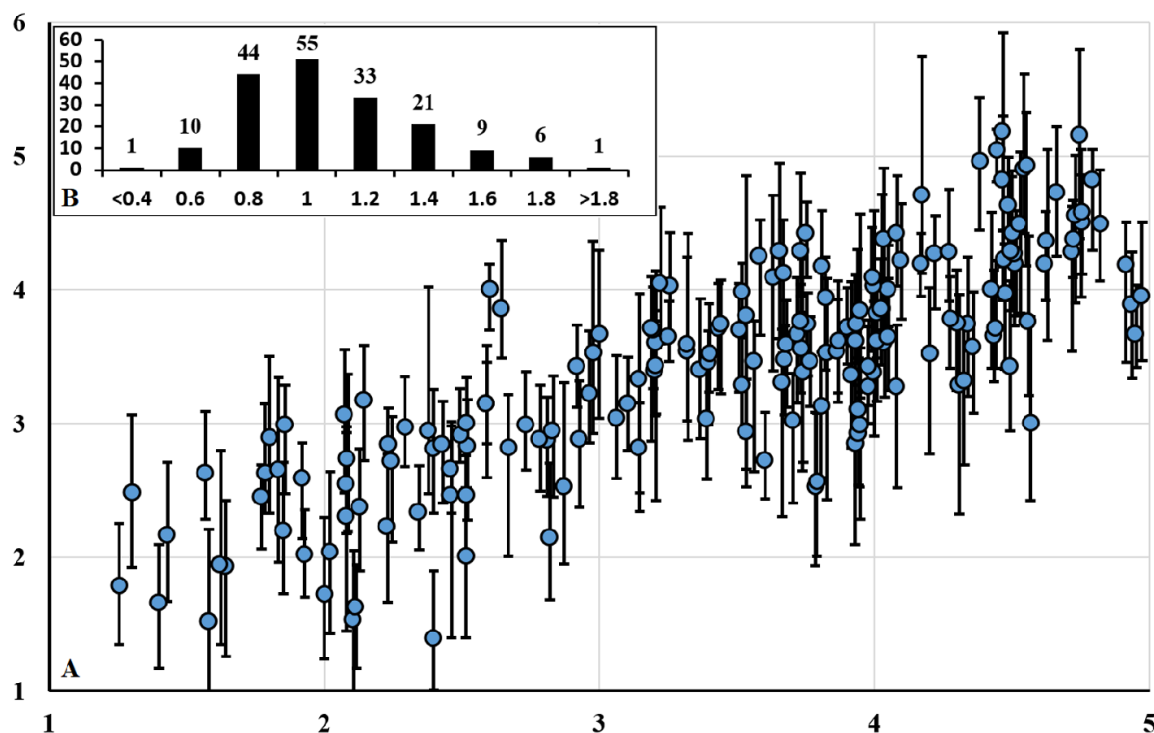


Figure 4. A. The comparison of experimentally determined and predicted $\lg(\text{IC}_{50})$ values of the complete dataset, including the complex variants obtained from molecular dynamics simulations (15 for each inhibitor/enzyme complex). Prediction was performed using equations #14.1–#14.5. In each case, we used the equation where that inhibitor was not used in the learning procedure. For each inhibitor, the mean and the range from minimum to maximum are shown. On the abscissas axis are the experimentally determined values of $\lg(\text{IC}_{50})$. On the vertical axis are the predicted values. B. The distribution of the predicted range for each inhibitor.

equations from the E1 set generally included only the magnitude of the change in the energy of covalent interactions in the ligand and receptor, for which there was no analog in E2.

The parameters from the E2 set were obtained by averaging the corresponding data calculated at a several regular step on the molecular dynamics simulation trajectory. The predicted value of $\lg(\text{IC}_{50})$ can be calculated for a variant of the enzyme-inhibitor complex at each such step (Fig. 4). In this case, the scatter of the predicted values can be interpreted as an additional quality measure (analogous to the scatter of data in experimental measurements). The larger the scatter of the predicted values, the more doubtful is the prediction result.

Finally, we note that by restricting to the similar spatial position of docking variants for structurally similar compounds and adding indicator variables related to the data source, it is possible to generate a set of equations that adequately predicts the IC_{50} value for inhibitors of SARS-CoV-2 M^{pro}.

FUNDING

The work was performed within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021–2030) (No. 122030100170-5).

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any research involving humans or using animals as objects.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

Supplementary materials are available in the electronic version at the journal site (pbmc.ibmc.msk.ru).

REFERENCES

1. Ivanova Ya.O., Voronina A.I., Skvortsov V.S. (2022) The prediction of SARS-CoV-2 main protease inhibition with filtering by position of ligand. *Biomeditsinskaya Khimiya*, **68**(6), 444–458.
2. Han S.H., Goins C.M., Arya T., Shin W.-J., Maw J., Hooper A., Sonawane D.P., Porter M.R., Bannister B.E., Crouch R.D., Lindsey A.A., Lakatos G., Martinez S.R., Alvarado J., Akers W.S., Wang N.S., Jung J.U., Macdonald J.D., Stauffer S.R. (2022) Structure-based optimization of ML300-derived, noncovalent inhibitors targeting the severe acute respiratory syndrome coronavirus 3CL protease (SARS-CoV-2 3CL(pro)). *J. Med. Chem.*, **65**(4), 2880–2904. DOI: 10.1021/acs.jmedchem.1c00598

3. Case D.A., Aktulga H.M., Belfon K., Ben-Shalom I.Y., Berryman J.T., Brozell S.R., Cerutti D.S., Cheatham T.E. III, Cisneros G.A., Cruzeiro V.W.D., Darden T.A., Forouzes N., Giambaeu G., Giese T., Gilson M.K., Gohlke H., Goetz A.W., Harris J., Izadi S., Izmailov S.A., Kasavajhala K., Kaymak M.C. et al (2023) Amber 2023, University of California, San Francisco.
4. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235-242. DOI: 10.1093/nar/28.1.235
5. Schrodinger (Schrodinger, LLC, New York, NY). Retrieved September 02, 2023 from <https://www.schrodinger.com/>
6. Harder E., Damm W., Maple J., Wu C., Reboul M., Xiang J.Y., Wang L., Lupyan D., Dahlgren M.K., Knight J.L., Kaus J.W., Cerutti D.S., Krilov G., Jorgensen W.L., Abel R., Friesner R.A. (2015) OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.*, **12**(1), 281-296. DOI: 10.1021/acs.jctc.5b00864
7. Mikurova A.V., Skvortsov V.S., Grigoryev V.V. (2020) Generalized predictive model of estimation of inhibition of muscarinic receptors M1-M5. *Biomedical Chemistry: Research and Methods*, **3**(3), e00129. DOI: 10.18097/bmcrm00129
8. Massova I., Kollman P.A. (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives Drug Discovery Design*, **18**, 113-135. DOI: 10.1023/A:1008763014207
9. SYBYL-X, Tripos, St. Louis, MO, USA.
10. Mikurova A.V., Skvortsov V.S. (2018) Prediction of progesterone affinity for the human progesterone receptor based on corrected RBA data. *Biomedical Chemistry: Research and Methods*, **1**(4), e00080. DOI: 10.18097/BMCRM00080
11. Gentile F., Fernandez M., Ban F., Ton A.-T., Mslati H., Perez C.F., Leblanc E., Yaacoub J.C., Gleave J., Stern A., Wong B., Jean F., Strynadka N., Cherkasov A. (2021) Automated discovery of noncovalent inhibitors of SARS-CoV-2 main protease by consensus deep docking of 40 billion small molecules. *Chemical Science*, **12**(48), 15960-15974. DOI: 10.1039/d1sc05579h
12. Deodato D., Asad N., Dore T.M. (2022) Discovery of 2-thiobenzimidazoles as noncovalent inhibitors of SARS-CoV-2 main protease. *Bioorganic Med. Chem. Lett.*, **72**, 128867. DOI: 10.1016/j.bmcl.2022.128867

Received: 22. 09. 2023.
 Revised: 18. 10. 2023.
 Accepted: 20. 10. 2023.

ПРЕДСКАЗАНИЕ ИНГИБИРОВАНИЯ ГЛАВНОЙ ПРОТЕАЗЫ SARS-CoV-2 НА МОДЕЛЯХ КОМПЛЕКСОВ ИНГИБИТОР-ФЕРМЕНТ

Я.О. Иванова*, В.С. Скворцов

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; *эл. почта: yana.emris@gmail.com

Проанализирован набор уравнений линейной регрессии, предсказывающих величину IC_{50} для 180 конкурентных ингибиторов главной протеазы SARS-CoV-2. Проведена симуляция молекулярной динамики комплексов фермент-ингибитор, либо имеющих известную кристаллическую структуру, либо промоделированных методом молекулярного докинга с наложенным ограничением на отбор конечных поз по сходству со структурными аналогами. В качестве независимых переменных использовали величины энергетических вкладов, полученных при расчёте двумя вариантами метода MMPBSA (MMGBSA), изменения свободной энергии комплекса, и ряд физико-химических характеристик ингибиторов. При обучении для подвыборок, полученных из различных источников, использовали индикаторные переменные, чтобы нивелировать имеющиеся систематические отклонения целевой величины. Качество предсказания оценивали по процедуре скользящего контроля методом выбрасывания по одному и по 20% выборки. Средняя ошибка при предсказании величины $\lg(IC_{50})$ составила 0,45 логарифмической единицы при общей ширине диапазона значений 3,71. Рассмотрена зависимость устойчивости предсказания от вариативности комплекса в процедуре молекулярной динамики.

Полный текст статьи на русском языке доступен на сайте журнала (<http://pbmc.ibmc.msk.ru>).

Ключевые слова: SARS-CoV-2; главная протеаза; конкурентные ингибиторы; QSAR

Финансирование. Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021-2030 годы) (№ 122030100170-5).

Поступила в редакцию: 22.09.2023; после доработки: 18.10.2023; принята к печати: 20.10.2023.