# Supporting Information for Developing a Machine Learning Interatomic Potential for Non-Covalent Interactions in Proteins

*Lejia Zeng[1,2], Xintong Zhang[1,2], Yuchan Pei[1,2], Lifeng Zhao[2], Lan Hua[2], Jincai Yang[2], Niu Huang[1,2,*]*

[1]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing 102206, China

[2]National Institute of Biological Sciences, 7 Science Park Road, Zhongguancun Life Science Park, Beijing 102206, China

E-mail: huangniu@nibs.ac.cn

# Contents

# 1 Dimer Extraction

Dimers were extracted from protein structures using an in-house script. For each type of fragment, a corresponding SMARTS[1]pattern was generated, and SMILES[2] patterns were also employed to define the three-dimensional templates. Substructure searches across all collected proteins were then carried out through SMARTS-based substructure matching implemented in RDKit[3] Python package. The coordinates of matching atoms of each hit were mapped onto the template atoms defined by corresponding SMILES patterns. When isolation of the fragment required cleavage of a C–C or C–N bond, the bond was cut at the appropriate position, the resulting terminal atom was first marked as a dummy atom, and then capped with hydrogen to maintain valency.

Once 17 chemical groups had been collected from a given protein, dimers were searched for by identifying fragment pairs in close spatial proximity. Each interacting pair was initially labeled as the shortest distance between 2 and 4 Å. And the residue in which two fragments are located should be separated by at least 2 amino acids to avoid bias from proximity effects.

# 2 Active Learning Workflow

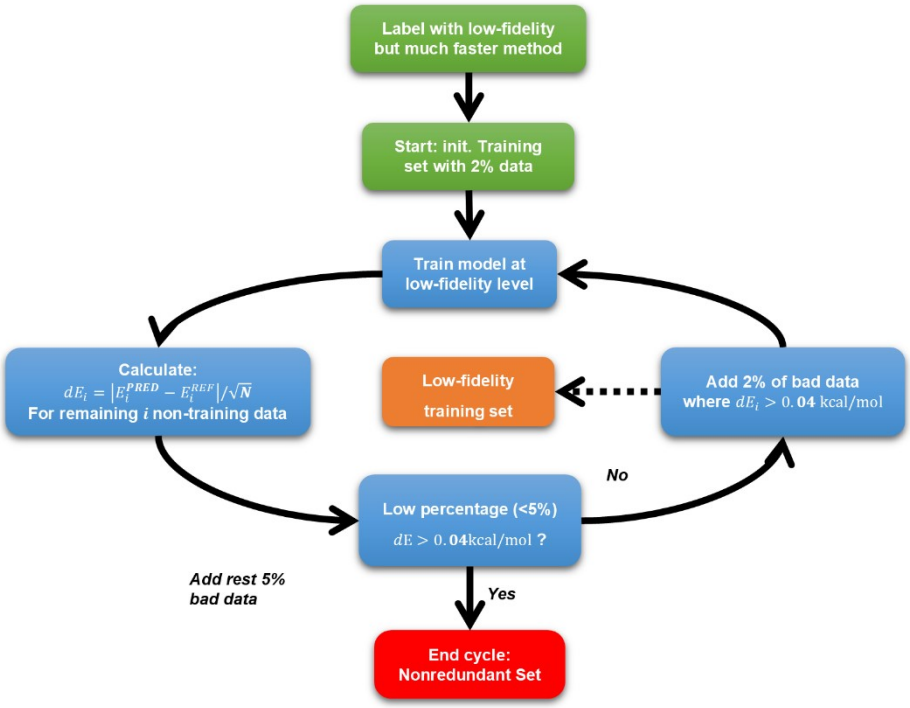


**Figure S1.** Active learning workflow for data redundancy.

# 3 Random Sampling

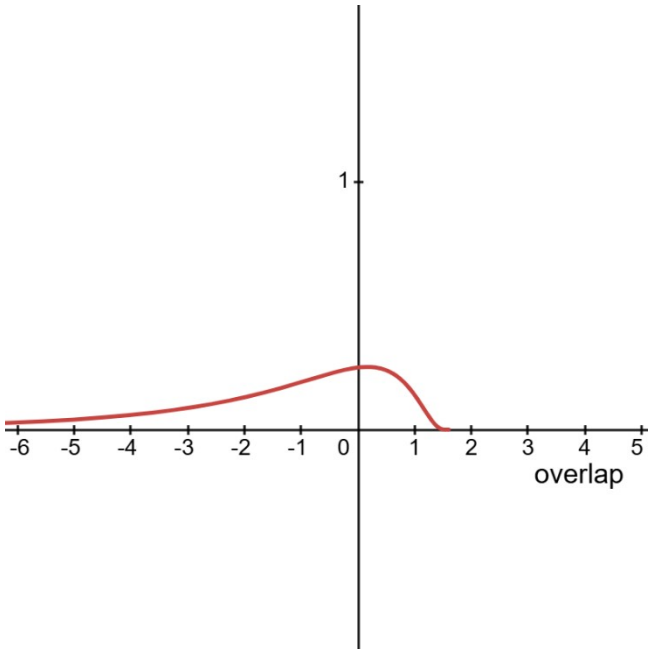


**Figure S2.** Acceptance probability curve for intermolecular overlap used in random sampling. Curve visualization generated via https://www.desmos.com/calculator/a8imk7uszg.

# 4 Model Performance on Benchmark Sets

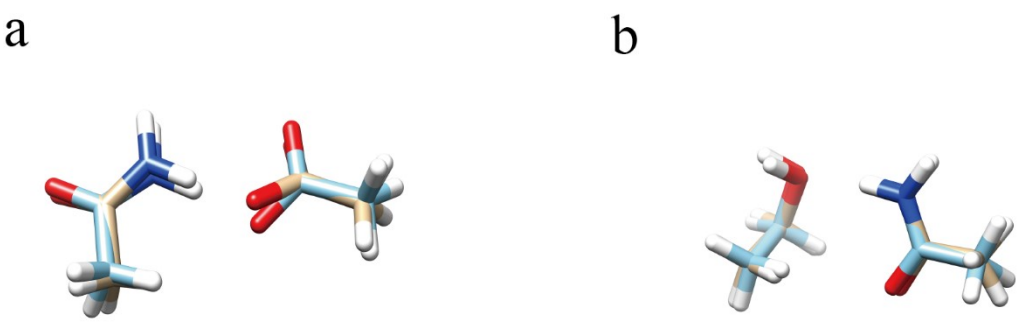


**Figure S3.** Conformational changes observed in post-optimization. Initial lowest-energy structures (tan) and their optimized counterparts (sky blue) exhibit minimal structural deviations, as quantified by root-mean-square deviation (RMSD) values of (a) 0.399 Å for ACEM-ACET and (b) 0.352 Å for ACEM-ETOH.

# 5 Comparison with ANI-2x

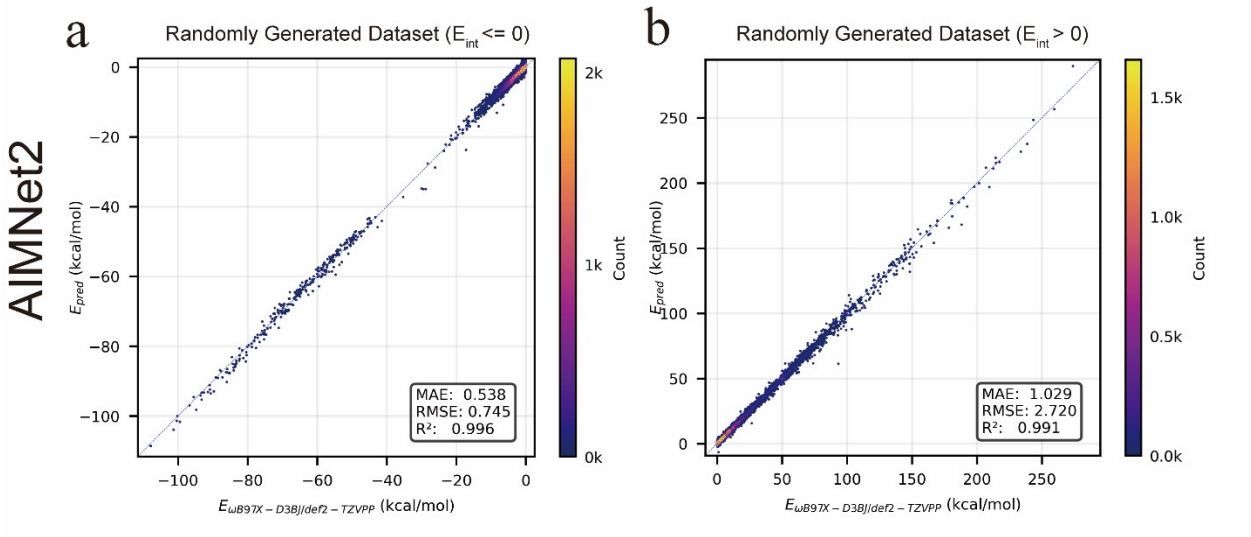


**Figure S4.** Interaction energies calculated at ωB97X-D3BJ/def2-TZVPP level and predicted by AIMNet2 for the randomly generated dataset with interaction energy below 0 kcal/mol (a), and above 0 kcal/mol (b).

# 6 Exploring NCI Patterns in PDB

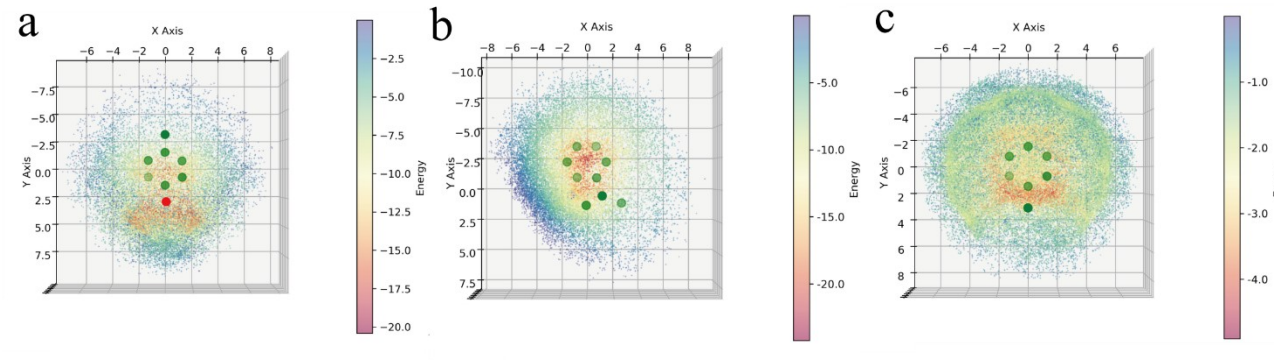


**Figure S5.** Bottom view of spatial distributions for ETAM-PMPO (a), ETAM-MIND (b), and MBZ-MSM (c). Each scatter point represents the position of the fragment's central atom (N or S) relative to the central fragment. The color of each scatter point corresponds to the energy scale shown in the adjacent color bar. Circles on the xy-plane represent heavy atoms of the central fragment: green for carbon atoms, red for oxygen atoms, and blue for nitrogen.

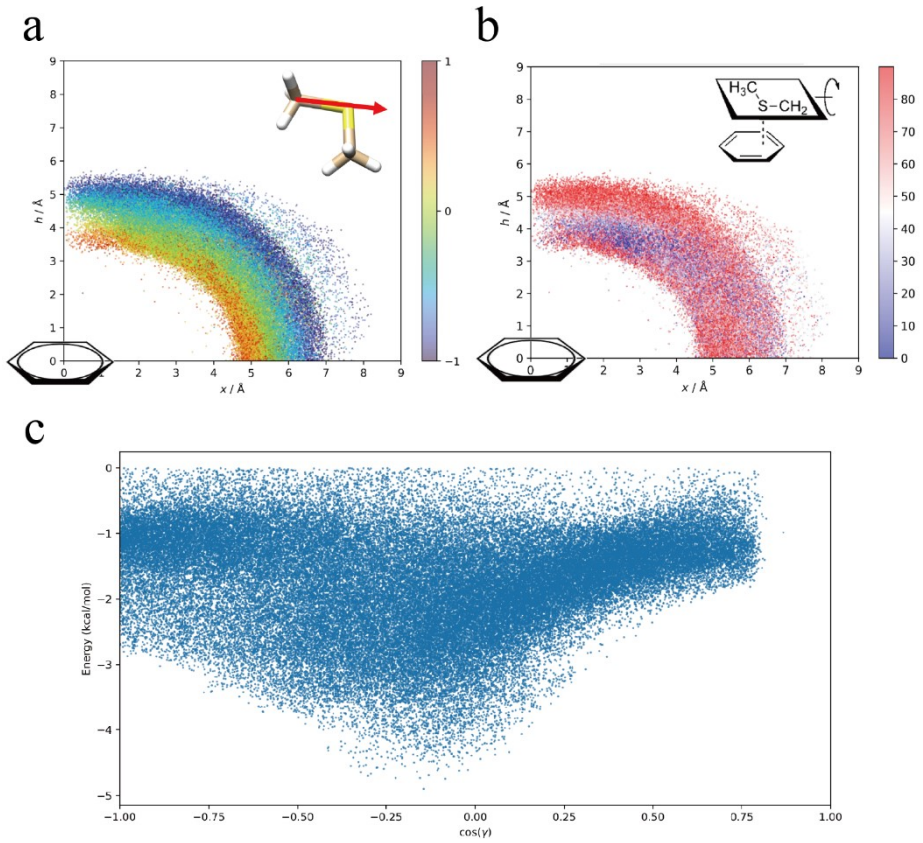


**Figure S6.** Radial distribution of MSM around aromatic rings, with $x$ as the centroid shift (Å) and $h$ as the height above the plane (Å), with point colors indicating the trend of the geometric parameter of interest. (a) Orientation distribution of the C→S vector. The color bar denotes orientation ($\cos\gamma$): values > 0 (red) correspond to the C→S vector pointing toward the ring center, whereas values < 0 (blue) indicate the opposite direction. The inset (upper right) is the schematic of the C–S bond definition. (b) Interplanar angle distribution between the plane of the MSM fragment and the aromatic ring plane. The inset (upper right) illustrates the definition of the MSM fragment plane used in the analysis. (c) Predicted interaction energies (kcal/mol) for MBZ-MSM as a function of the C–S bond orientation ($\cos\gamma$, where 1 = toward the aromatic ring, –1 = away, 0 = parallel).

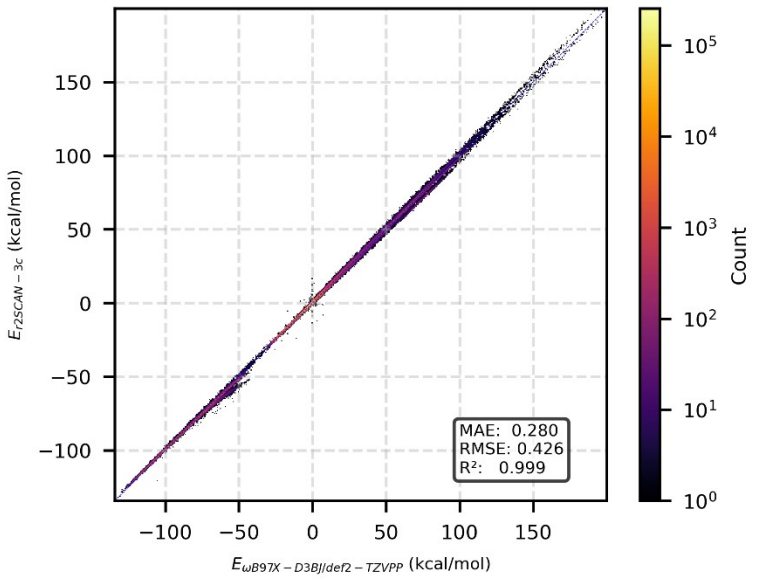


**Figure S7.** Comparison of noncovalent interaction energies calculated at the ωB97X-D3BJ/def2-TZVPP and r²SCAN-3c levels for all dimers in the PDB-FRAGID dataset.

# 7 NCI Visualization

The Reduced Density Gradient (RDG) method[4] is used to visualize weak interactions. In this study, weak interaction analysis[5] was performed using the Multiwfn program[6], and the corresponding visualizations were generated with VMD software[7]. In the weak interaction analysis, the gradient isosurfaces of the RDG function are typically colored on a scale ranging from blue-green to red, corresponding to interaction strength from strong attraction to strong repulsion. Blue regions indicate strong attractive interactions, green regions correspond to weak van der Waals interactions, and red regions denote strong steric repulsion.

# 8 Supporting Information Tables

**Table S1.** Data volume of each fragment

| Fragment type | Total number* | Training number**(%)*** |
|---|---|---|
| ACEM | 4,830,133 | 533,043 (11.03) |
| ACET | 5,246,390 | 1,170,051 (22.30) |
| ETAM | 1,480,001 | 347,779 (23.50) |
| ETOH | 3,299,189 | 374,098 (11.34) |
| ETSH | 350,411 | 79,419 (22.66) |
| MBZ | 3,380,598 | 252,517 (7.47) |
| MGDM | 2,520,733 | 474,908 (18.84) |
| Imidazole | 1,377,566 | 284,607 (20.66) |
| MIND | 1,325,598 | 194,941 (14.71) |
| PMPO | 1,697,412 | 258,379 (15.22) |
| MSM | 949,756 | 141,498 (14.90) |
| N1PA | 1,103,276 | 179,516 (16.27) |
| NMA | 22,026,093 | 876,434 (3.98) |
| PRPA | 9,306,132 | 527,630 (5.67) |
| HOH | 3,561,146 | 327,658 (9.20) |

*The total number of occurrences in the original dataset.

**The number of occurrences in the training set.

***Percentage of total data used for training.

**Table S2.** Calculation time of each fragment

| Fragment Type | Time*(in hours) | | |
|---|---|---|---|
| | ωB97X-D3BJ/def2-TZVPP | r²SCAN-3c | PANIP |
| ACEM | 36 days, 20:03:14.2 | 5:58:16.3 | 0:39:18.0 |
| ACET | 20 days, 9:32:51.3 | 5:48:37.5 | 0:37:32.9 |
| ETAM | 45 days, 11:24:32.5 | 5:14.52.6 | 0:40:42.3 |
| ETOH | 34 days, 18:38:02.8 | 4:53.28.3 | 0:39:19.3 |
| ETSH | 29 days, 0:16:31.1 | 5:04:10.7 | 0:39:03.9 |
| MBZ | 53 days, 17:01:56.3 | 8:05:08.1 | 0:44:10.8 |
| MGDM | 66 days, 5:37:29.9 | 7:05:37.8 | 0:41:39.7 |
| MIMD | 56 days, 10:43:06.7 | 7:28:09.5 | 0:40:31.7 |
| MIME | 54 days, 12:43:22.4 | 7:22:51.6 | 0:41:42.3 |
| MIMM | 58 days, 13:51:29.8 | 7:14:30.6 | 0:40:60.0 |
| MIND | 108 days, 13:58:41.7 | 11:36:17.0 | 0:46:57.3 |
| PMPO | 80 days, 17:58:59.7 | 9:09:32.6 | 0:44:01.9 |
| MSM | 35 days, 0:20:00.7 | 4:58:28.8 | 0:39:58.2 |
| N1PA | 94 days, 2:16:00.9 | 10:53:37.0 | 0:46:23.8 |
| NMA | 47 days, 5:44:10.5 | 6:27:22.1 | 0:42:11.5 |
| PRPA | 36 days, 13:47:55.2 | 5:08:38.6 | 0:41:15.4 |
| HOH | 12 days, 21:28:45.8 | 3:09:31.4 | 0:35:22.9 |
| Total | 463 days, 11:19:06.1 | 4days, 19:39:10.6 | 6:17:55.6 |

*All benchmark calculations were performed on the computer cluster with an Intel Xeon Platinum 8368 CPU @ 2.60 GHz with 504 GB RAM.

**Table S3.** Calculation time of benchmark datasets

| Datasets | Time* (in minutes) | |
|---|---|---|
| | PANIP | AIMNet2 |
| Low-Energy Dataset | 00:14.9 | 00:21.9 |
| Optimized Low-Energy Dataset | 00:09.4 | 00:14.5 |
| CSD | 09:52.9 | 12:10.9 |
| Randomly Generated Dataset | 04:37.3 | 06:11.5 |
| Total | 14:54.5 | 18:58.8 |

*All benchmark calculations were performed on the computer cluster with an NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM, CUDA Version 12.9.

**Table S4.** Comparison of RMSE and MAE values for MLIP predictions across datasets curated from the intermolecular interaction benchmark database[8] used in this study*.

| Dataset | Size | PANIP | | | AIMNet2 | | |
|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ |
| **Charged** | | | | | | | |
| DES370K-AIMNet2-elements | 15069 | 0.35 | 0.46 | 0.99 | 0.56 | 0.82 | 0.98 |
| Ionic_H_bonds | 41 | 0.38 | 0.46 | 0.98 | 0.66 | 0.80 | 0.94 |
| NCIA_IHB100x10 | 64 | 0.51 | 0.59 | 0.99 | 1.12 | 1.74 | 0.95 |
| NENCI-2021-AIMNet2-elements | 1719 | 0.36 | 0.47 | 0.99 | 0.75 | 1.03 | 0.98 |
| **Neutral** | | | | | | | |
| A24 | 1 | 0.04 | 0.04 | N/A** | 0.38 | 0.38 | N/A |
| HB104 | 12 | 0.55 | 0.62 | 0.79 | 0.75 | 0.90 | 0.56 |
| NCIA_D442x10-AIMNet2-elements | 380 | 0.46 | 0.54 | 0.87 | 16.45 | 43.36 | -855.01 |
| NCIA_HB300SPXx10 | 179 | 0.43 | 0.51 | 0.92 | 1.11 | 1.28 | 0.49 |
| NCIA_HB375x10 | 884 | 0.43 | 0.51 | 0.96 | 0.65 | 0.89 | 0.88 |
| NCIA_R739x5-AIMNet2-elements | 11 | 0.45 | 0.48 | 0.91 | 1.16 | 1.54 | 0.10 |
| R160x6 | 113 | 0.36 | 0.44 | 0.94 | 0.69 | 1.04 | 0.66 |
| S66a8 | 314 | 0.36 | 0.45 | 0.96 | 0.59 | 0.72 | 0.89 |
| S66x8 | 305 | 0.36 | 0.45 | 0.97 | 0.65 | 0.83 | 0.91 |
| sulfurx8 | 91 | 0.32 | 0.40 | 0.95 | 0.86 | 1.15 | 0.56 |

*RMSE and MAE values are reported in units of kcal/mol.

** For n = 1, $R^2$ is statistically meaningless owing to insufficient degrees of freedom and is therefore designated as “N/A”.

REFERENCES

(1) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2023-05-23).

(2) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules | Journal of Chemical Information and Modeling. https://pubs.acs.org/doi/10.1021/ci00057a005 (accessed 2023-05-23).

(3) RDKit. http://www.rdkit.org/ (accessed 2023-05-23).

(4) Johnson, E. R.; Keinan, S.; Mori-Sánchez, P.; Contreras-García, J.; Cohen, A. J.; Yang, W. Revealing Noncovalent Interactions. J. Am. Chem. Soc. 2010, 132 (18), 6498–6506. https://doi.org/10.1021/ja100936w.

(5) Lu, T.; Chen, Q. Visualization Analysis of Weak Interactions in Chemical Systems. In Comprehensive Computational Chemistry (First Edition); Yáñez, M., Boyd, R. J., Eds.; Elsevier: Oxford, 2024; pp 240–264. https://doi.org/10.1016/B978-0-12-821978-2.00076-3.

(6) Lu, T.; Chen, F. Multiwfn: A Multifunctional Wavefunction Analyzer. Journal of Computational Chemistry 2012, 33 (5), 580–592. https://doi.org/10.1002/jcc.22885.

(7) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. Journal of Molecular Graphics 1996, 14 (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(8) Nayal, K. S.; Cho, I.; Isayev, O. Benchmarking Universal Machine-Learned Interatomic Potentials for Intermolecular and Noncovalent Interactions. ChemRxiv 2026 (0218). https://doi.org/10.26434/chemrxiv.15000203/v1.