A POCKET-CENTRIC TRAIN-TEST SPLIT REDUCES DATA LEAKAGE IN STRUCTURE-BASED DEEP LEARNING

Sergei Evteev

medchem

Team



Yan Ivanenkov



Sergei Evteev



Alexey Ereshchenko



Alexander Malyshev

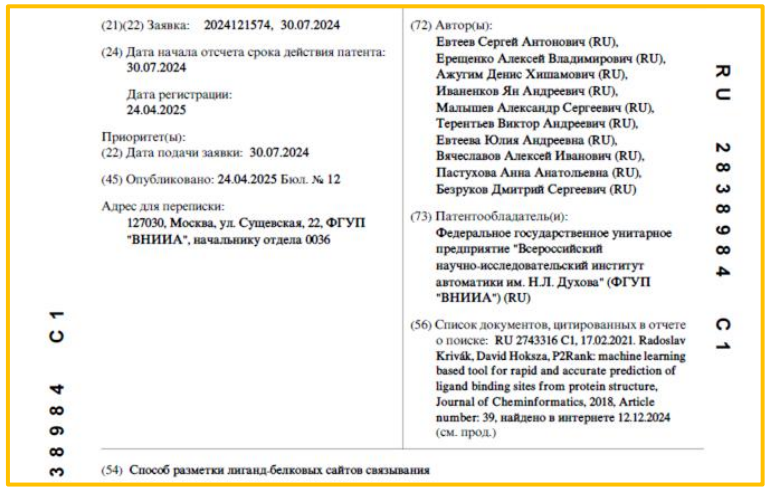


Anastasia Korzhenevskaya



Adjugim Denis

- Design of new bioactive compounds
- Hit-to-lead and lead optimization
- Software development for needs of medicinal chemistry
- Machine learning and language models in chemistry



RU 2838984 C1

(21)(22) Заявка: 2024121574, 30.07.2024

(24) Дата начала отсчета срока действия патента: 30.07.2024

Дата регистрации: 24.04.2025

Приоритет(ы): (22) Дата подачи заявки: 30.07.2024

(45) Опубликовано: 24.04.2025 Бюл. № 12

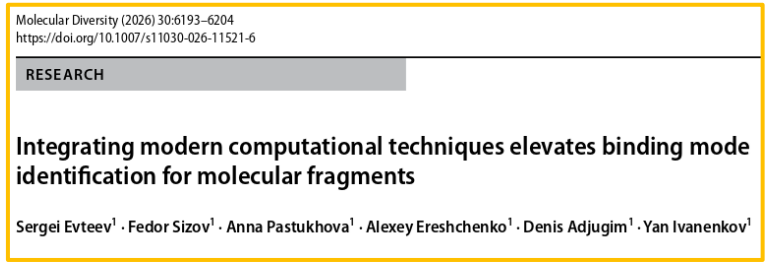
Адрес для переписки: 127030, Москва, ул. Сущевская, 22, ФГУП "ВНИИА", начальнику отдела 0036

(72) Автор(ы): Евтеев Сергей Антонович (RU), Ерещенко Алексей Владимирович (RU), Ажугим Денис Хишамович (RU), Иваненков Ян Андреевич (RU), Малышев Александр Сергеевич (RU), Терентьев Виктор Андреевич (RU), Евтеева Юлия Андреевна (RU), Вячеславов Алексей Иванович (RU), Пастухова Анна Анатольевна (RU), Безруков Дмитрий Сергеевич (RU)

(73) Патентообладатель(и): Федеральное государственное унитарное предприятие "Всероссийский научно-исследовательский институт автоматики им. Н.Л. Духова" (ФГУП "ВНИИА") (RU)

(56) Список документов, цитированных в отчете о поиске: RU 2743316 C1, 17.02.2021. Radoslav Krivák, David Hoksza, P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure, Journal of Cheminformatics, 2018, Article number: 39, найдено в интернете 12.12.2024 (см. прод.)

(54) Способ разметки лиганд-белковых сайтов связывания

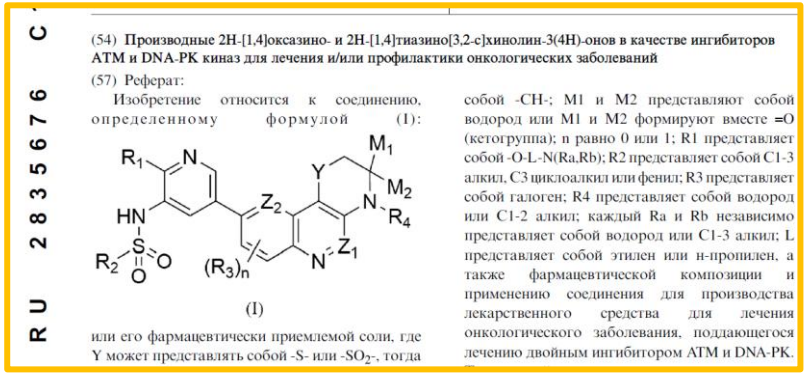


Molecular Diversity (2026) 30:6193–6204
https://doi.org/10.1007/s11030-026-11521-6

RESEARCH

Integrating modern computational techniques elevates binding mode identification for molecular fragments

Sergei Evteev[1] · Fedor Sizov[1] · Anna Pastukhova[1] · Alexey Ereshchenko[1] · Denis Adjugim[1] · Yan Ivanenkov[1]



RU 2835676 C1

(54) Производные 2H-[1,4]оксазино- и 2H-[1,4]тиазино[3,2-c]хинолин-3(4H)-онов в качестве ингибиторов ATM и DNA-PK киназ для лечения и/или профилактики онкологических заболеваний

(57) Реферат:

Изобретение относится к соединению, определенному формулой (I):

(I)

или его фармацевтически приемлемой соли, где Y может представлять собой -S- или -SO_2-, тогда собой -CH-; M1 и M2 представляют собой водород или M1 и M2 формируют вместе =O (кетогруппа); n равно 0 или 1; R1 представляет собой -O-L-N(Ra,Rb); R2 представляет собой C1-3 алкил, C3 циклоалкил или фенил; R3 представляет собой галоген; R4 представляет собой водород или C1-2 алкил; каждый Ra и Rb независимо представляет собой водород или C1-3 алкил; L представляет собой этилен или н-пропилен, а также фармацевтической композиции и применению соединения для производства лекарственного средства для лечения онкологического заболевания, поддающегося лечению двойным ингибитором ATM и DNA-PK.



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein–Ligand-Binding Sites

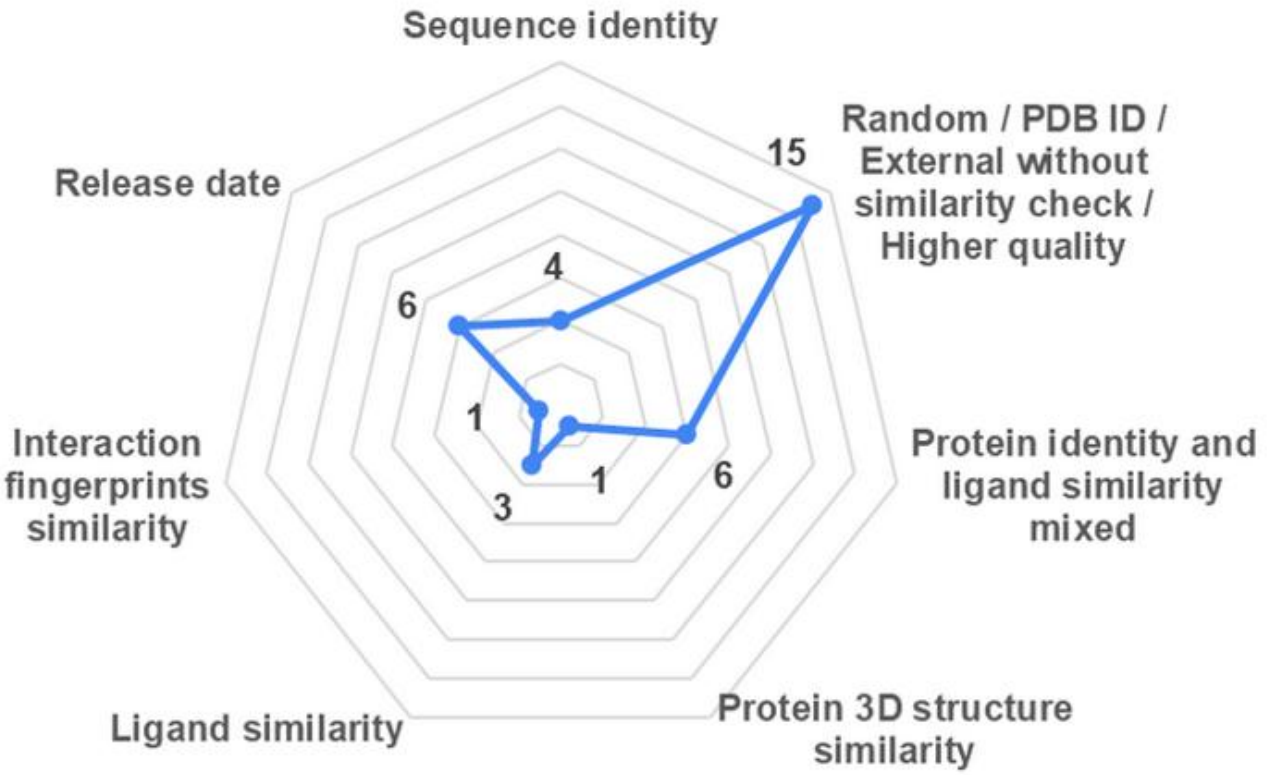
Sergei A. Evteev,* Alexey V. Ereshchenko, and Yan A. Ivanenkov

Crisis of train-test split in structure-based machine learning

- We analyzed **30** scientific articles describing new structure-based ML solutions published in **2023-2025**

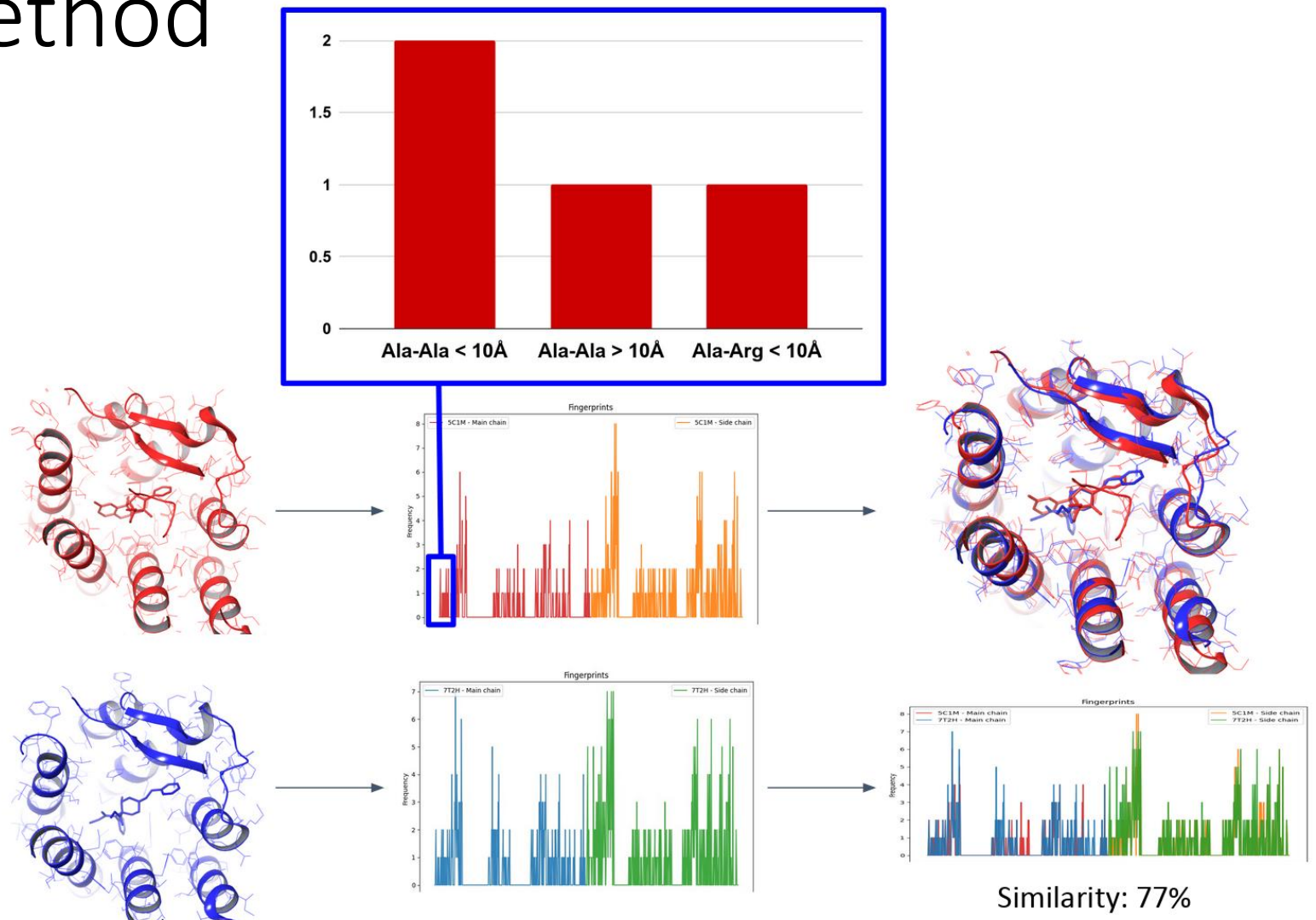
- Several most frequent train-test split techniques were revealed including random, release date, ligand-, sequence-based split

- Half of analyzed articles did not use any rational train test split such as sequence- or ligand-based!



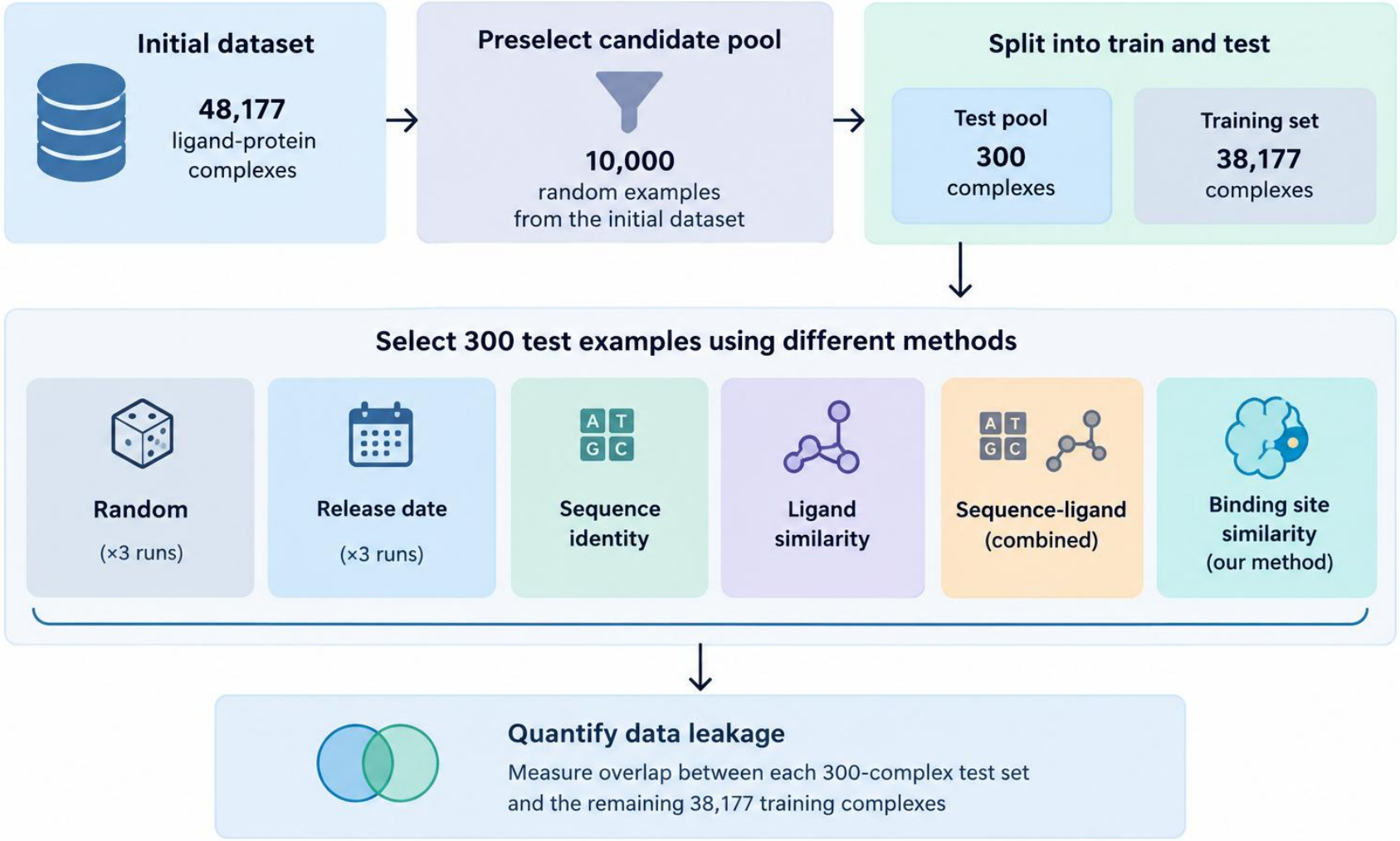
Site similarity method

- Our method represents 3D binding site geometry in 1D vector
- Each value represents frequency of two residue types found on distance less of more than 10 Å
- Separate analysis of side and main chains produces vector of 840 values



Experiment design

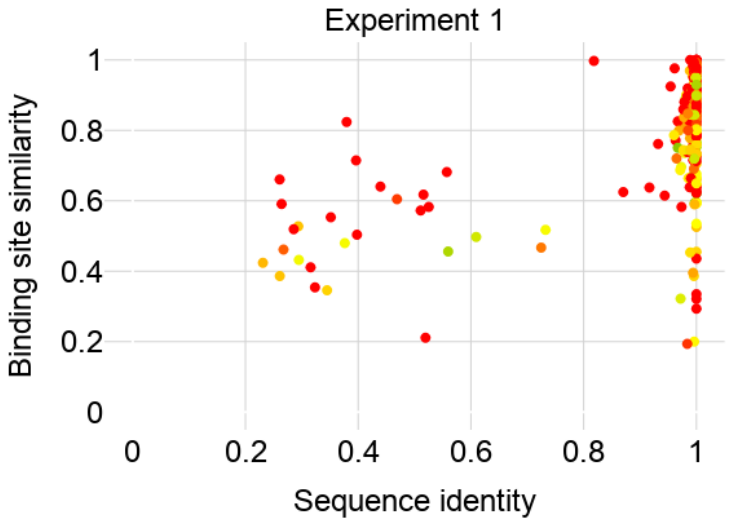
- Database of 48,000+ complexes was collected and split randomly into 38,000+ “train” and 10,000 preliminary “test” set
- Set of 300 test compounds was selected from preliminary test by each method
- Metrics were calculated and representative examples were analyzed

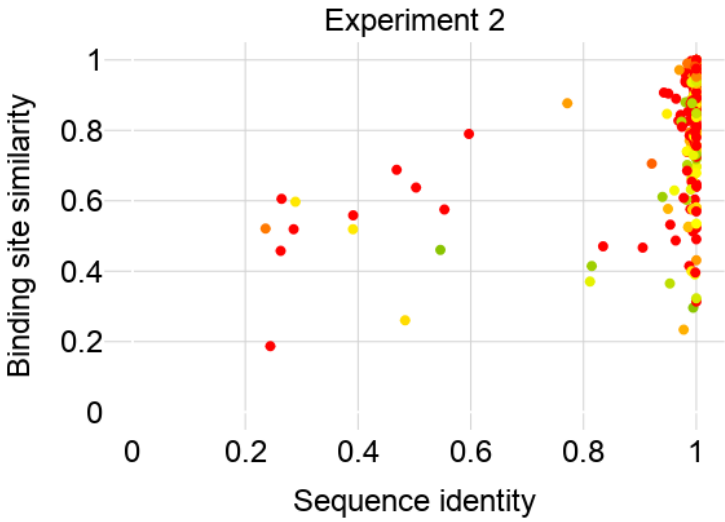


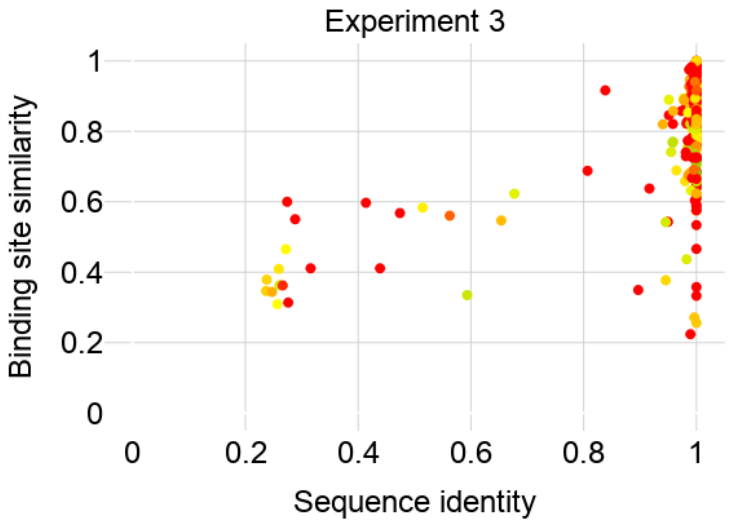
Random split

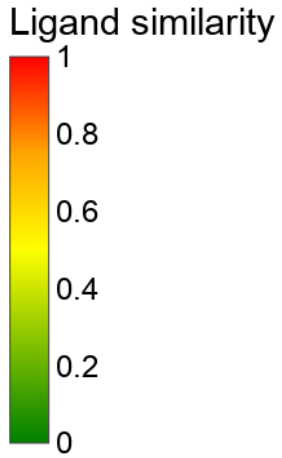
- Analysis revealed average values of 95.3% for sequence identity, 81.1% for site similarity, and 78.6% for ligand similarity

- Random split showed significant data leakage, demonstrating need for more rigorous strategies





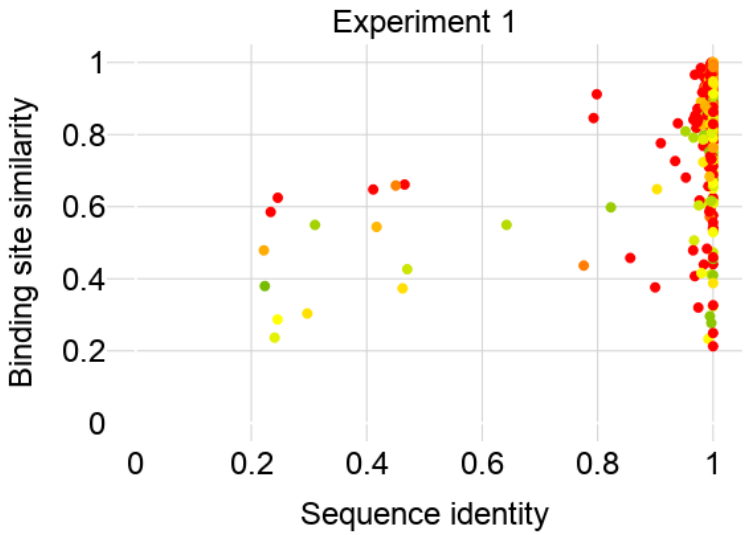


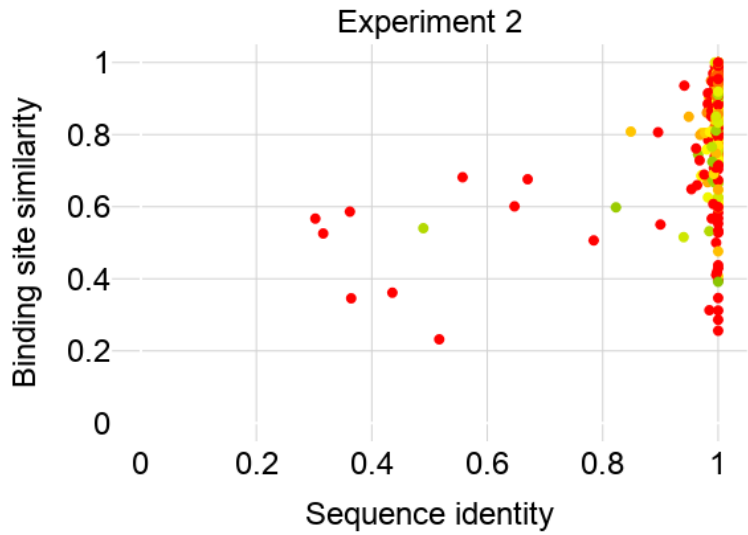


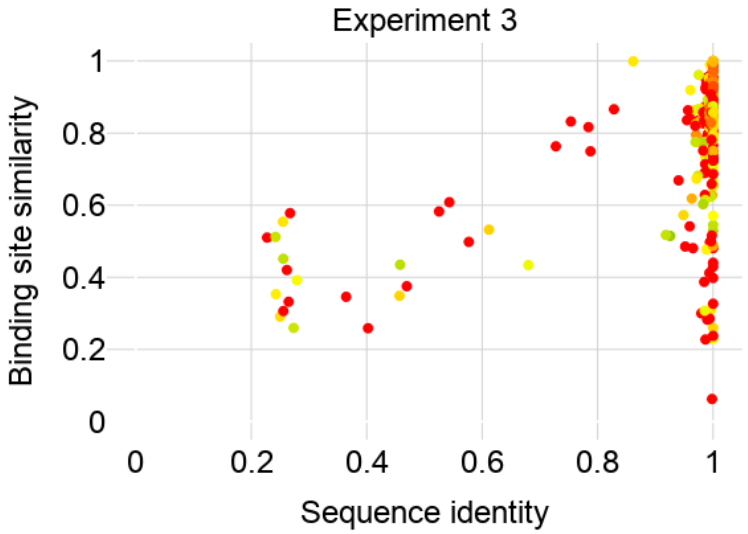
Release date split

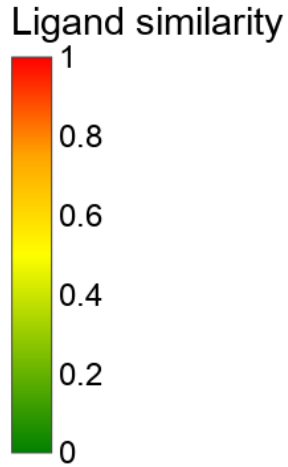
- Release data split led to same results as for random split with comparable values of average sequence identity (96.0% vs. 95.5%), site similarity (79.6% vs. 81.1%), and ligand similarity (76.5% vs. 78.6%)

- Splitting datasets by release date offers no significant improvement over random splitting



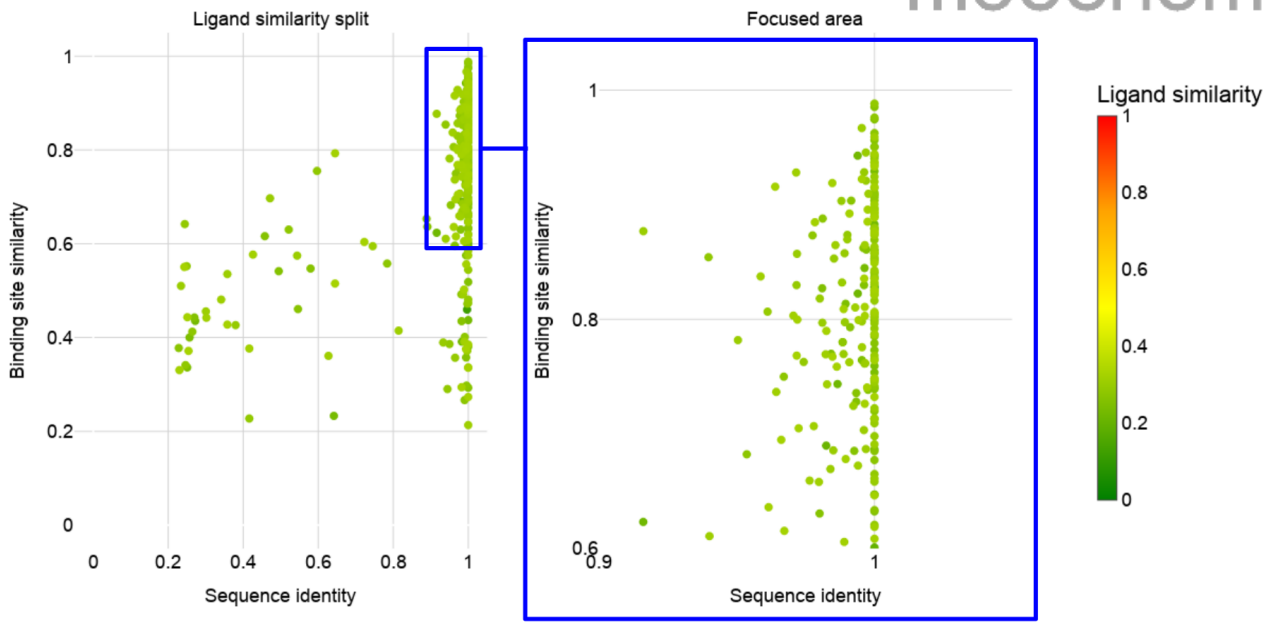


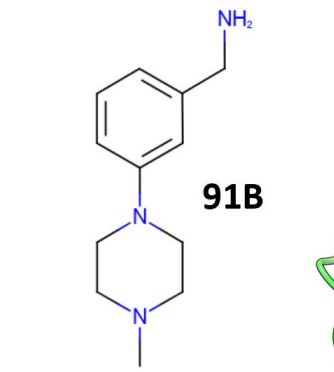


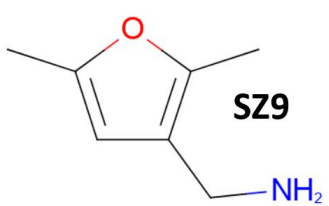


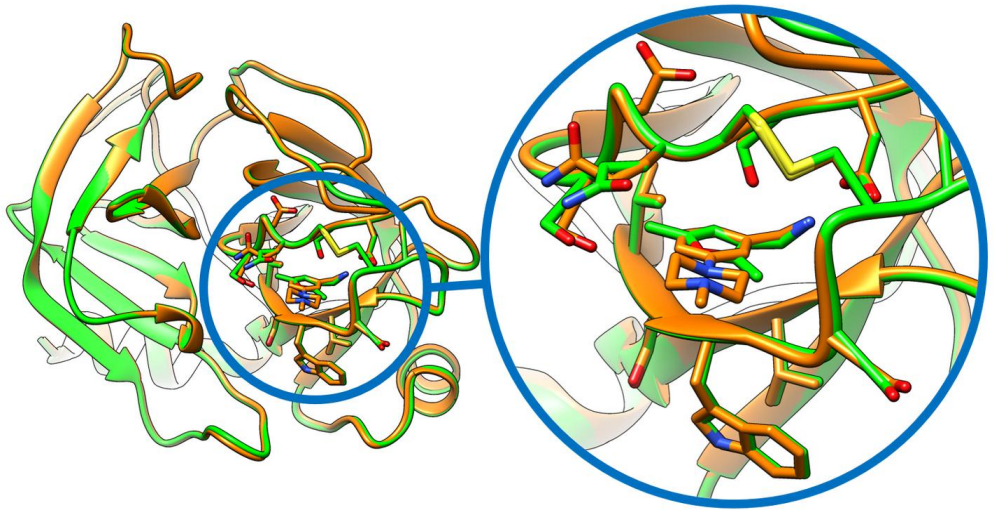
Ligand-based split

- Average ligand similarity was significantly decreased (28.7% vs. 78.6%)
- Ligand-based strategy did not change average sequence identity (91.8% vs. 95.5%) and site similarity (71.6% vs. 81.1%)
- As demonstrated by example of trypsin inhibitors, low ligand similarity (19.5%) does not mean low site similarity (98.8%)





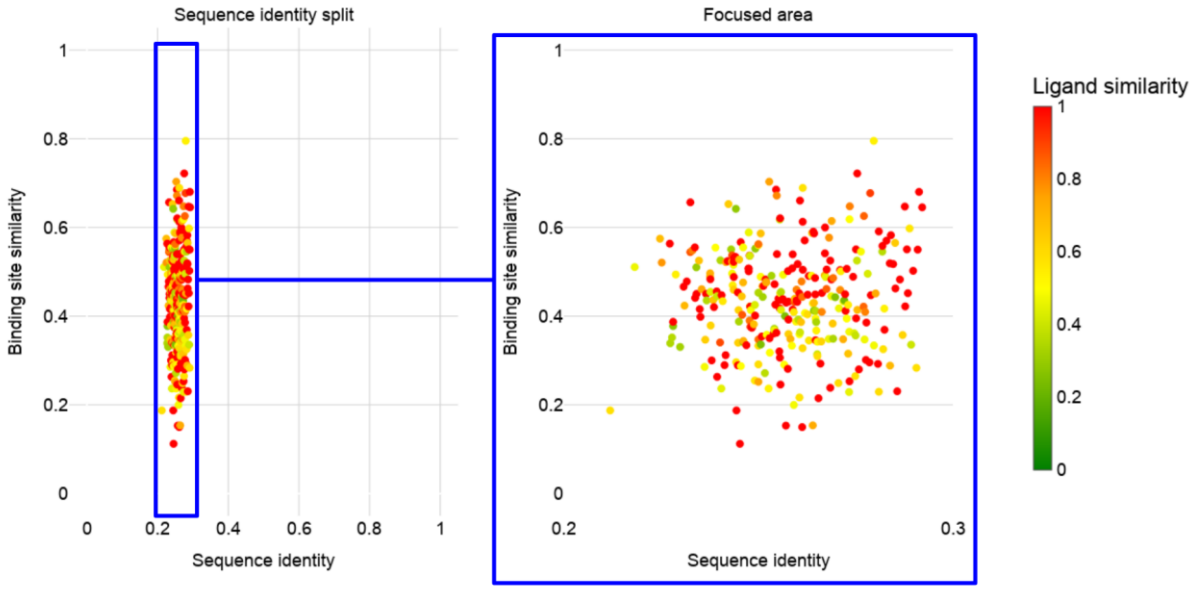




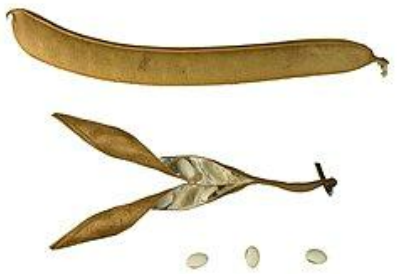
Sequence-based split

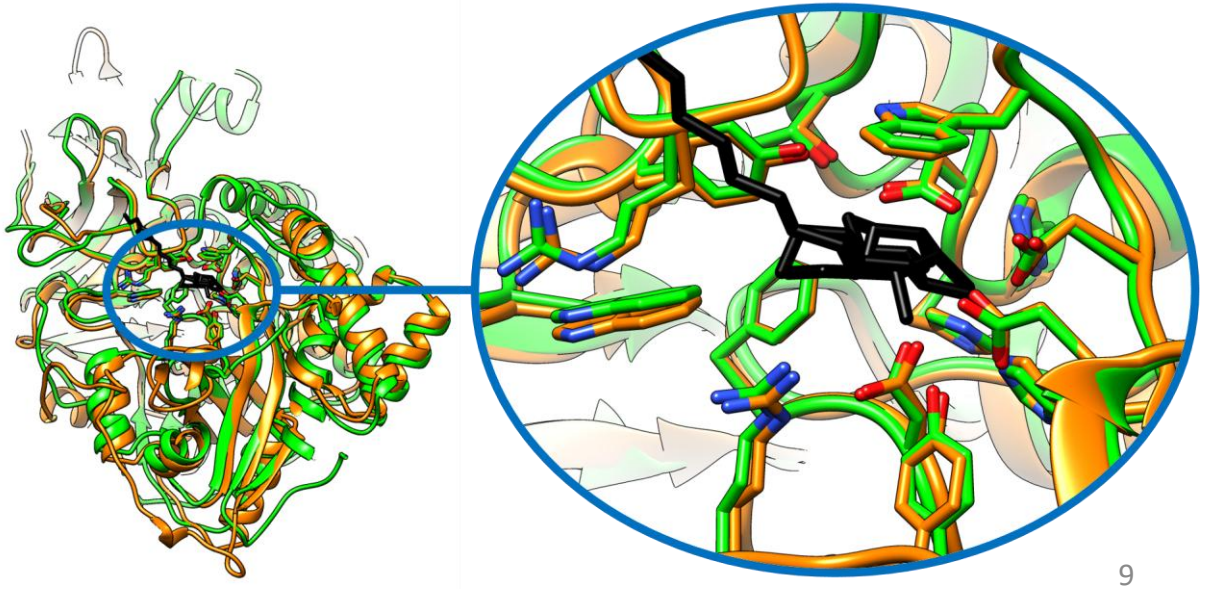
- Sequence-based approach decreased average sequence identity (25.8% vs. 95.5%) and site similarity (43.7% vs. 81.1%), but showed almost the same ligand similarity (74.9% vs. 78.6%)

- For example, low sequence identity (27%) between alpha-mannosidase of fruit fly and jack bean was accompanied by high site similarity (79.5%)





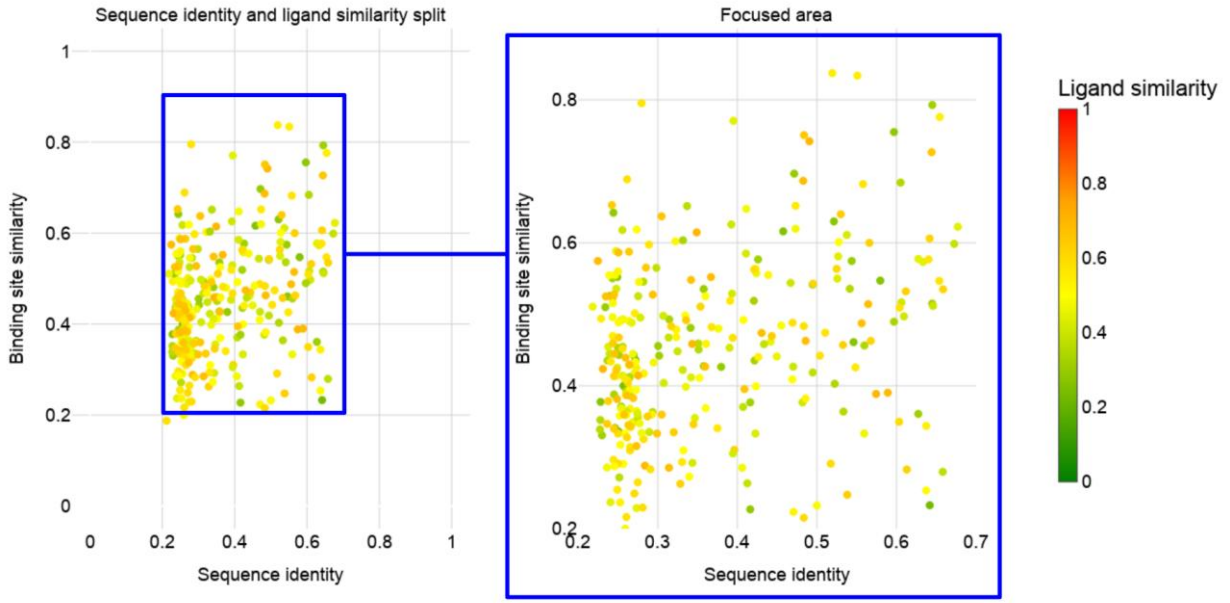


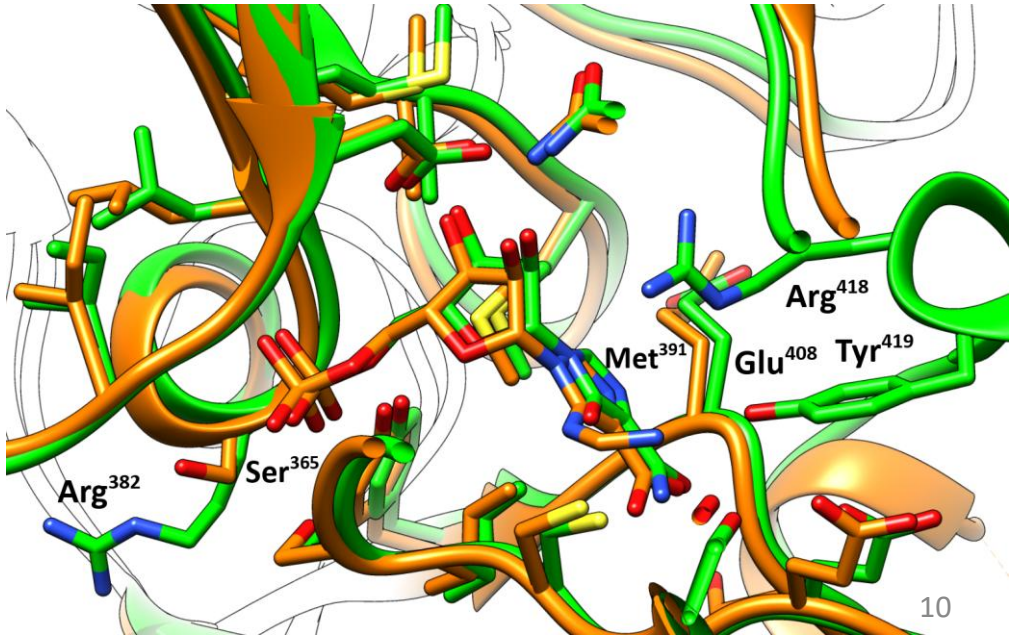


Split based on both sequence and ligand

- Mixed approach demonstrated improved values of average sequence identity (36.6% vs. 95.5%), site similarity (45.2% vs. 81.1%), and ligand similarity (49.5% vs. 78.6%)

- However, low sequence identity (36.5%) and ligand similarity (39.1%) between inosine-5`-monophosphate dehydrogenase of *Bacillus anthracis* and *Tritrichomonas suis* was accompanied by high site similarity (77.1%)

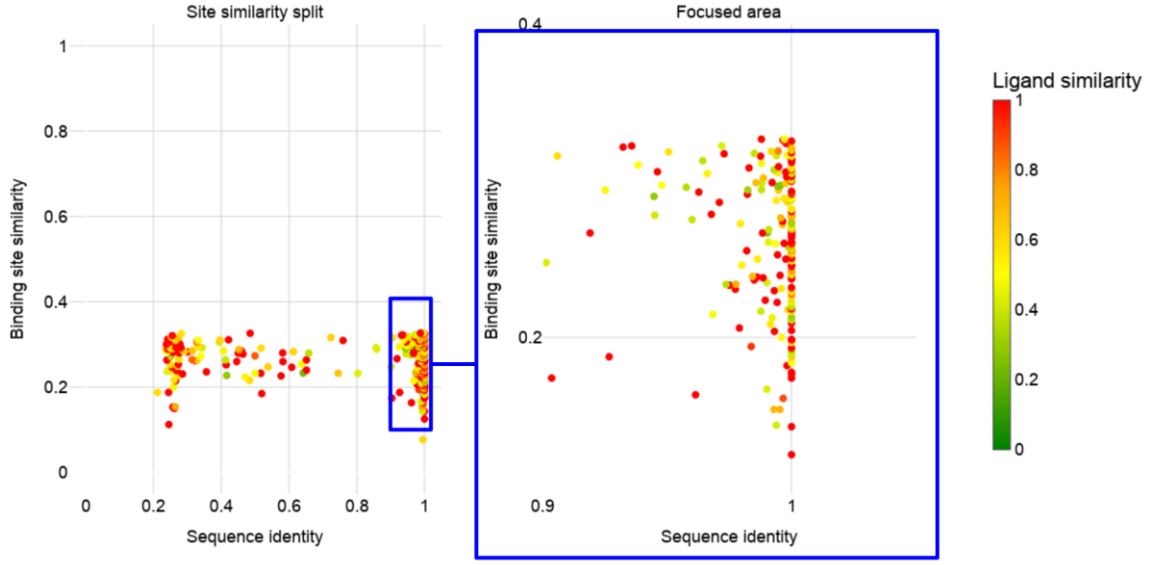


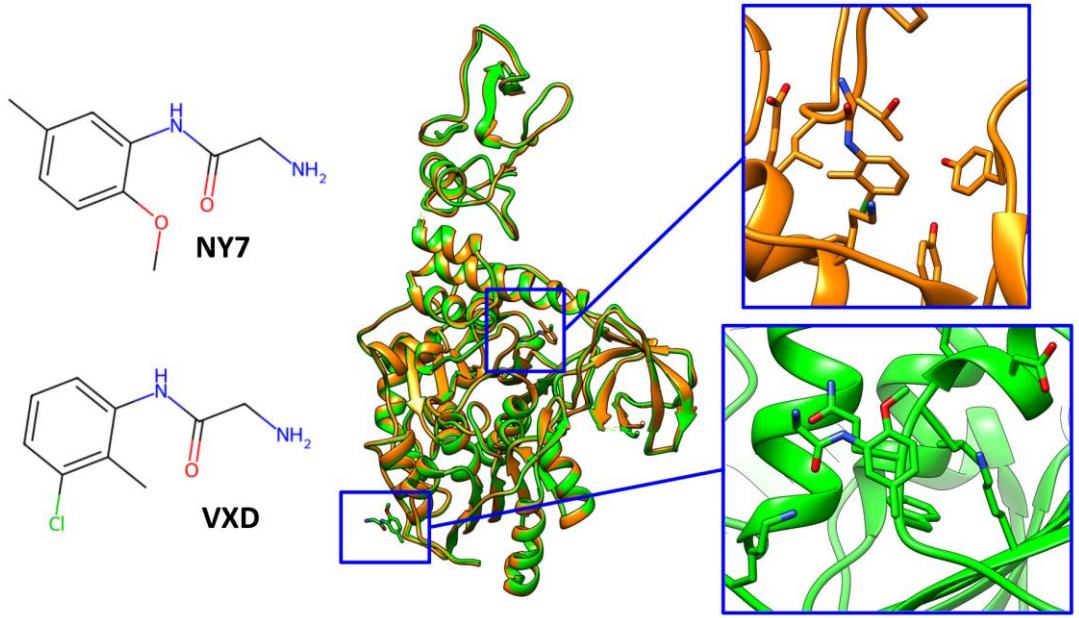


Site similarity split

- Site-similarity split improved average site similarity (26.1% vs. 81.1%) with minor change in sequence identity (79.6% vs. 95.5%) and ligand similarity (72.9% vs. 78.6%)

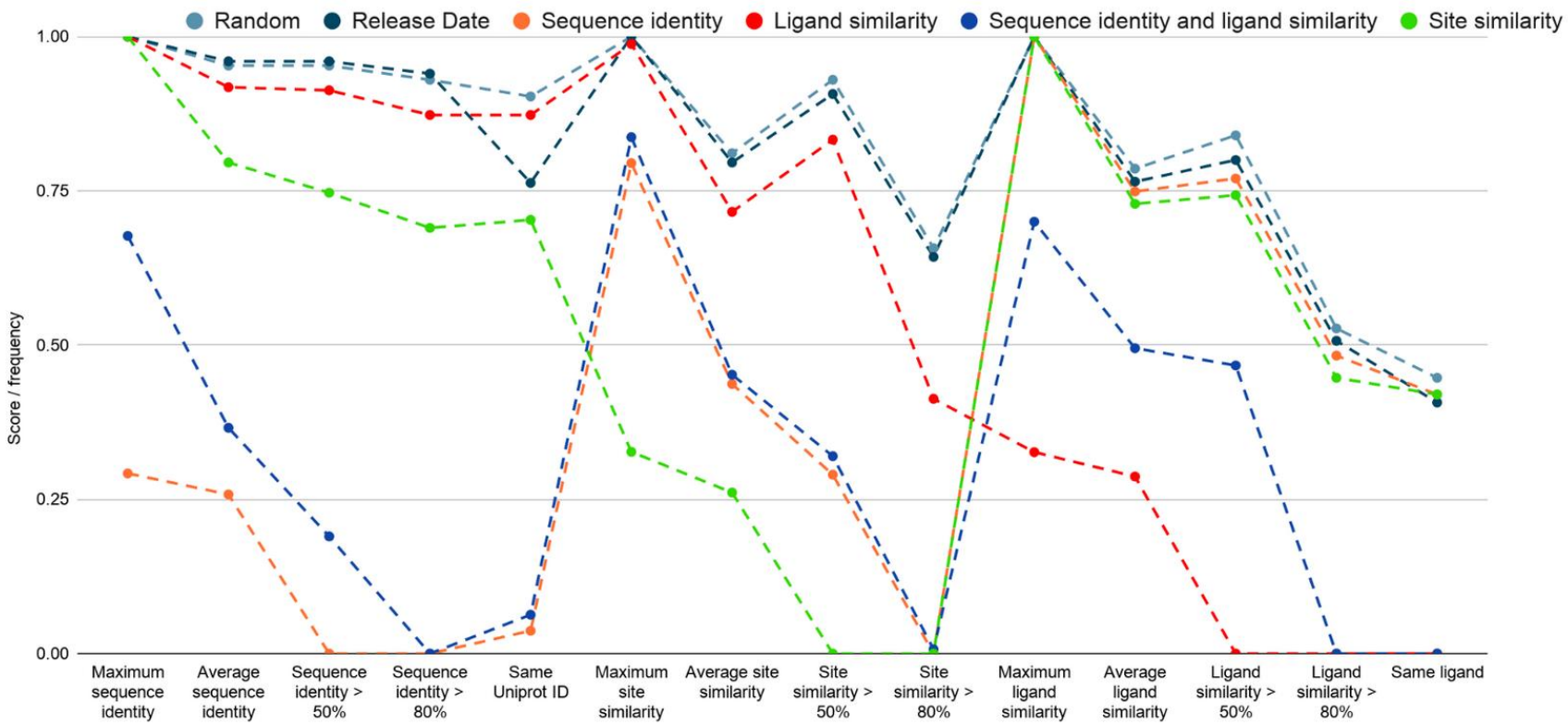
- Indeed, low site similarity (3%) revealed two distinct binding modes of SARS-CoV-2 helicase inhibitors despite high sequence identity (100%) and moderate ligand similarity (44.2%)





Key takeaways

- Random, release date, crystal quality split or unvetted use of external benchmarks leads to significant data leakage
- Ligand similarity split almost does not improve situation
- Sequence identity itself or in combination with ligand similarity is the best solution among earlier developed
- New site similarity technique was developed which provides robust and explainable foundation for train-test splitting



Thank you

Sergei Evteev
evteevsa1320@gmail.com
MED CHEM Ltd.

