THE USE OF ACTIVE LEARNING FOR EFFECTIVE EXPLORATION OF CHEMICAL UNIVERSE

ARTEM CHERKASOV
UBC

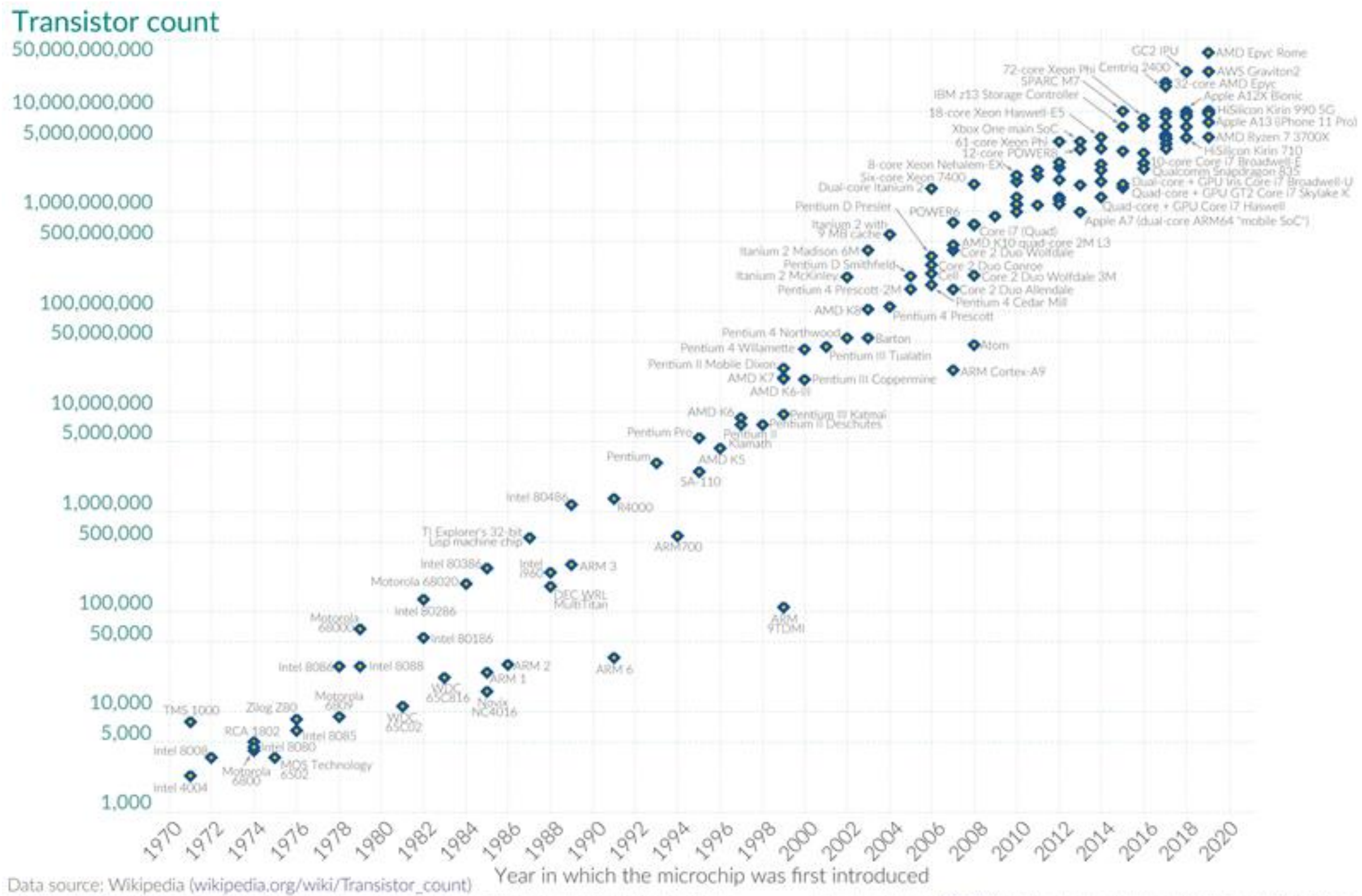
ICANN , SEPT 19, 2024

COMMERCIAL INTERESTS

ABT Therapeutics
LAST Innovation
Variational AI
OPTIC
PHOTONIC
RAKOVINA THERAPEUTICS
NIDO PHARMCAEUTICALS
ASTRA ZENECA

MOORE LAW : COMPUTERS BECOME CHEAPER AND MORE POWERFUL

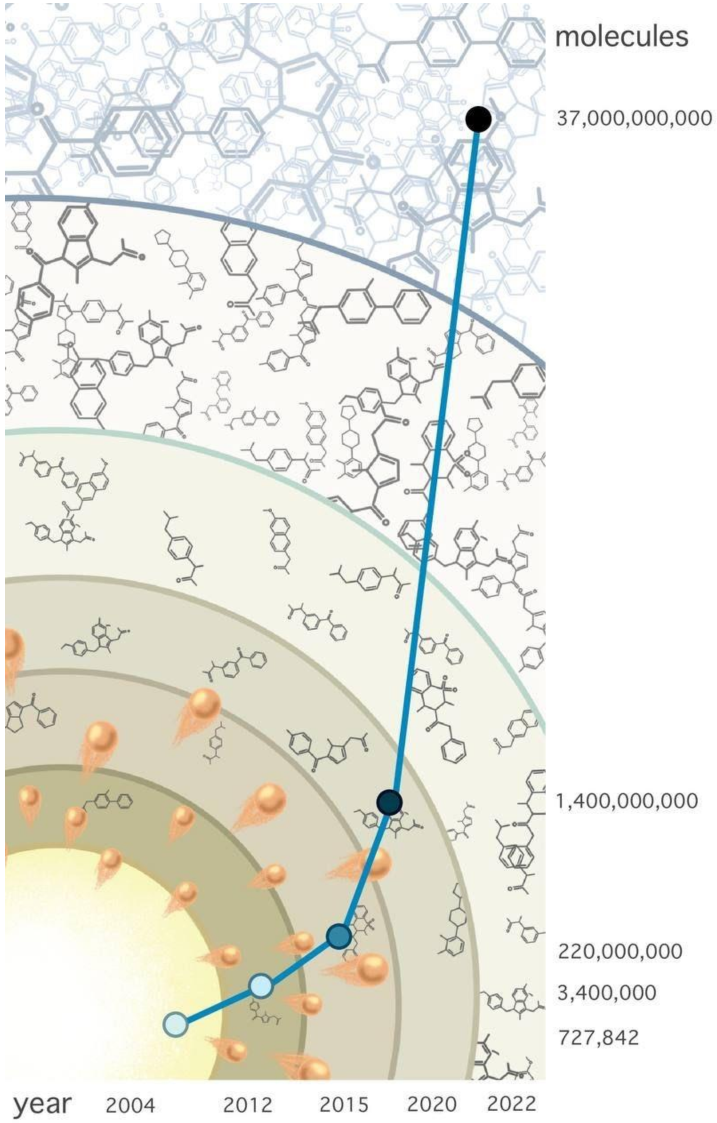
Example: number of transistors per chip multiplies by 2 every 2 years



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

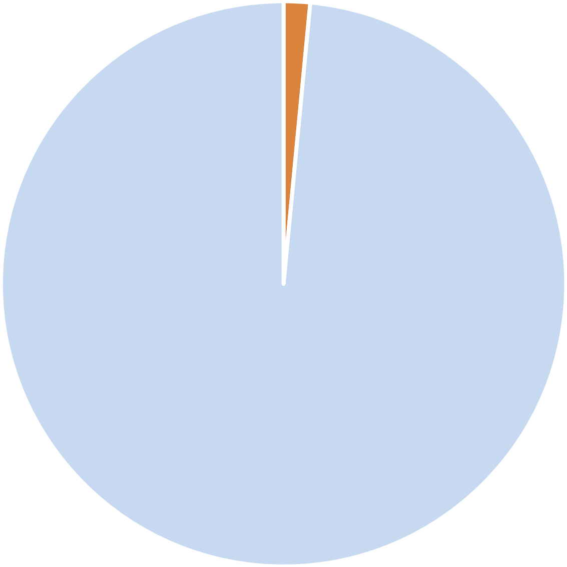
CHEMICAL SPACE REMAINS INACCESSIBLE TO DRUG DISCOVERY



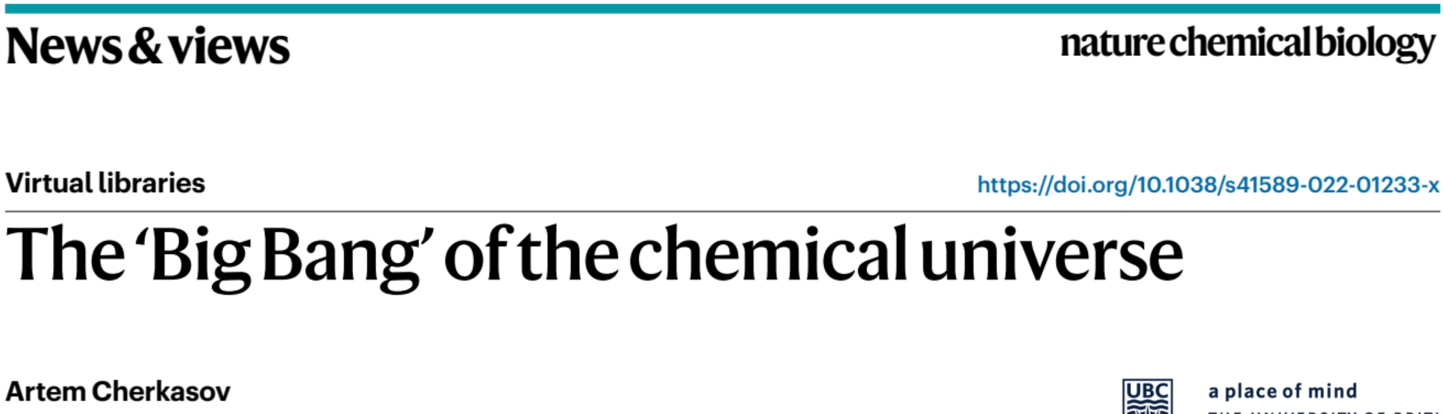
DOCKING MISSES OUT 99.9% OF ALREADY AVAILABLE MOLECULES

TOTAL NUMBER OF POSSIBLE DRUG-LIKE MOLECULES : 10^{60} - 10^{100}

10M DOCKING TIME 14 DAYS



1B MOLECULES DOCKING TIME 2.5YRS



News & views

nature chemical biology

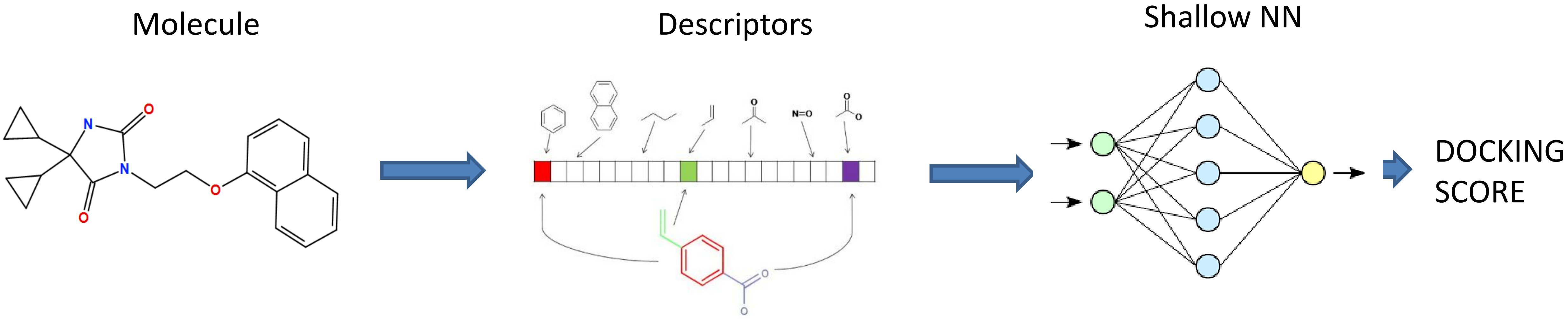
Virtual libraries

https://doi.org/10.1038/s41589-022-01233-x

The 'Big Bang' of the chemical universe

Artem Cherkasov

WHAT IF WE EMULATE DOCKING SCORES??

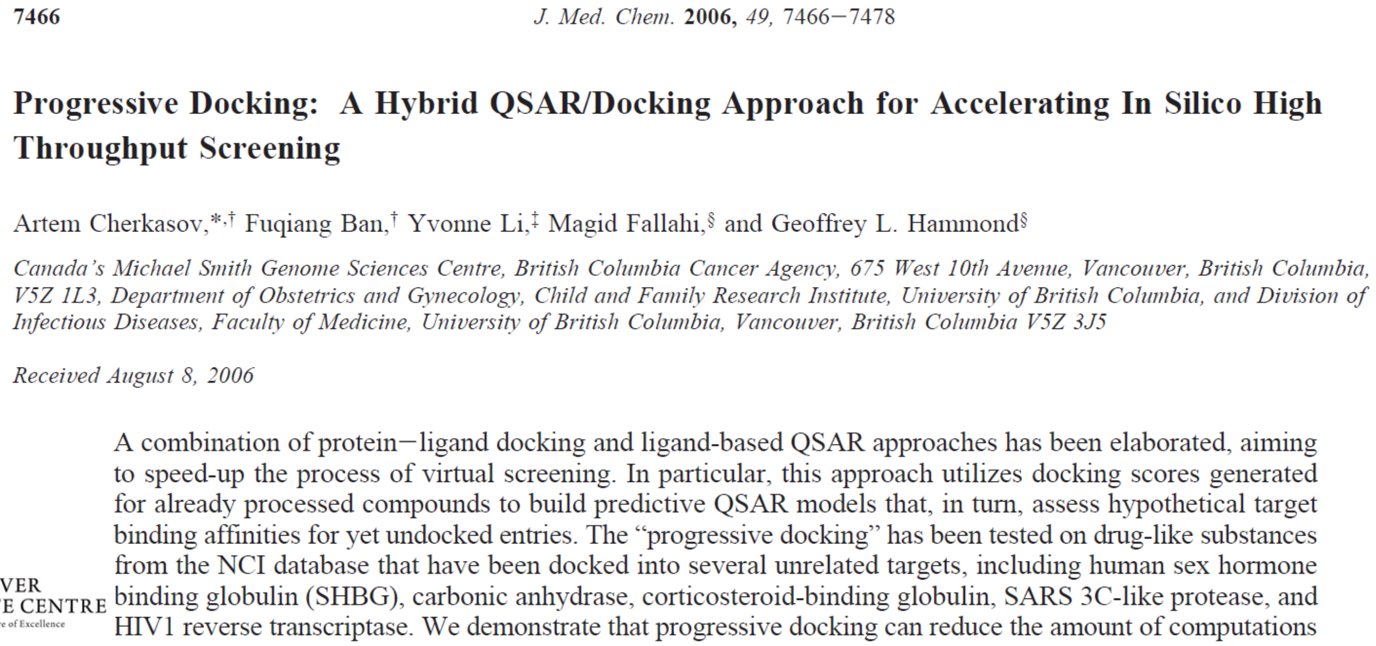


Molecule

Descriptors

Shallow NN

DOCKING SCORE



7466 J. Med. Chem. **2006,** *49,* 7466−7478

Progressive Docking: A Hybrid QSAR/Docking Approach for Accelerating In Silico High Throughput Screening

Artem Cherkasov,*[,†] Fuqiang Ban,[†] Yvonne Li,[‡] Magid Fallahi,[§] and Geoffrey L. Hammond[§]

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Department of Obstetrics and Gynecology, Child and Family Research Institute, University of British Columbia, and Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia V5Z 3J5

Received August 8, 2006

A combination of protein−ligand docking and ligand-based QSAR approaches has been elaborated, aiming to speed-up the process of virtual screening. In particular, this approach utilizes docking scores generated for already processed compounds to build predictive QSAR models that, in turn, assess hypothetical target binding affinities for yet undocked entries. The "progressive docking" has been tested on drug-like substances from the NCI database that have been docked into several unrelated targets, including human sex hormone binding globulin (SHBG), carbonic anhydrase, corticosteroid-binding globulin, SARS 3C-like protease, and HIV1 reverse transcriptase. We demonstrate that progressive docking can reduce the amount of computations

WHAT IF WE PREDICT DOCKING SCORES (AGAIN)??

MODELS TESTED:

DEEP NEURAL NETWORK (DNN)

RANDOM FOREST (RF)

SUPPORT VECTOR MACHINE (SVM)

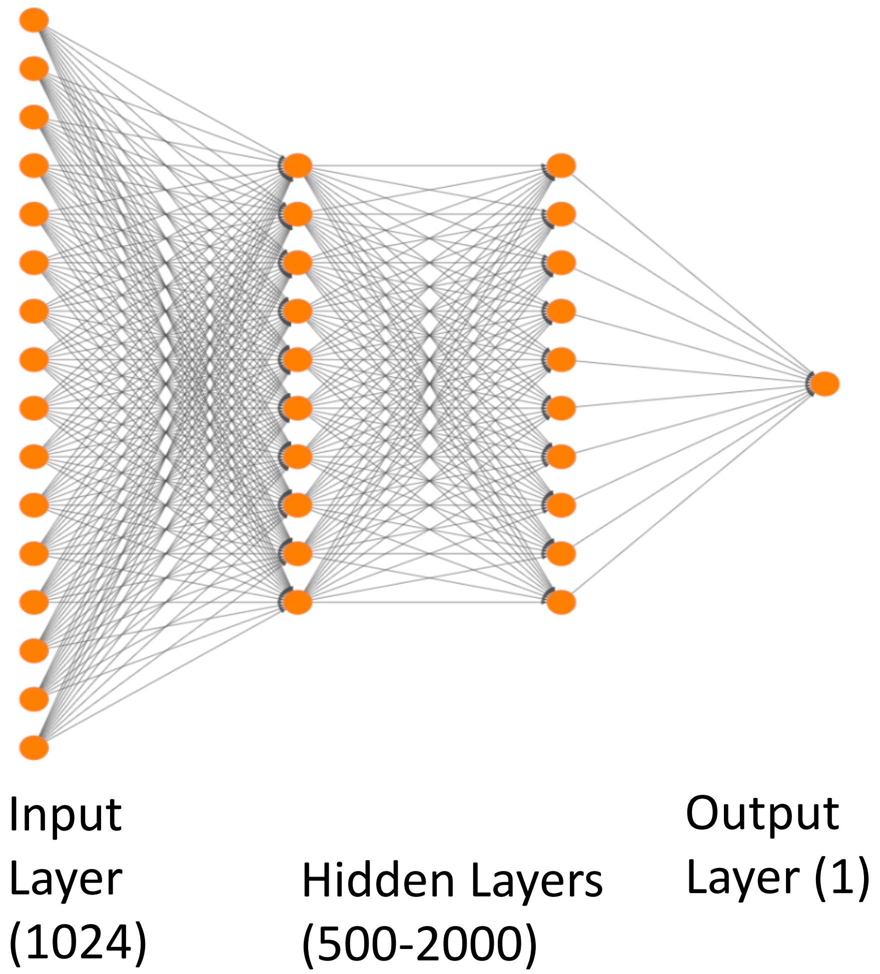
LOGISTIC REGRESSION (LR)

FINGERPRINTS TESTED

MACCS (166 BITS)

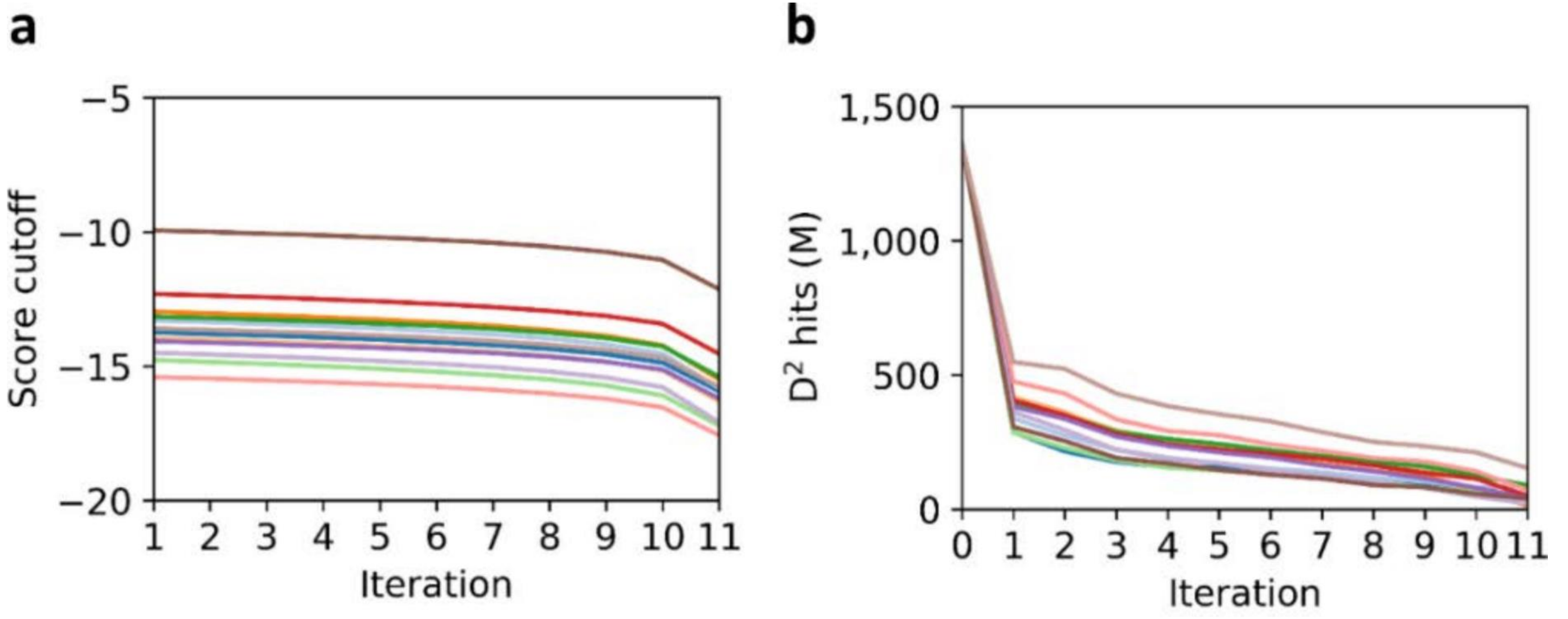
MORGAN WITH DIFFERENT R

PHARMACOPHORE



MORGAN WITH RADIUS 2 AND 1024 BITS + DNN SHOWED THE BEST PERFORMANCE

DEEP DOCKING PERFORMANCE ON 12 MAJOR DRUG TARGETS

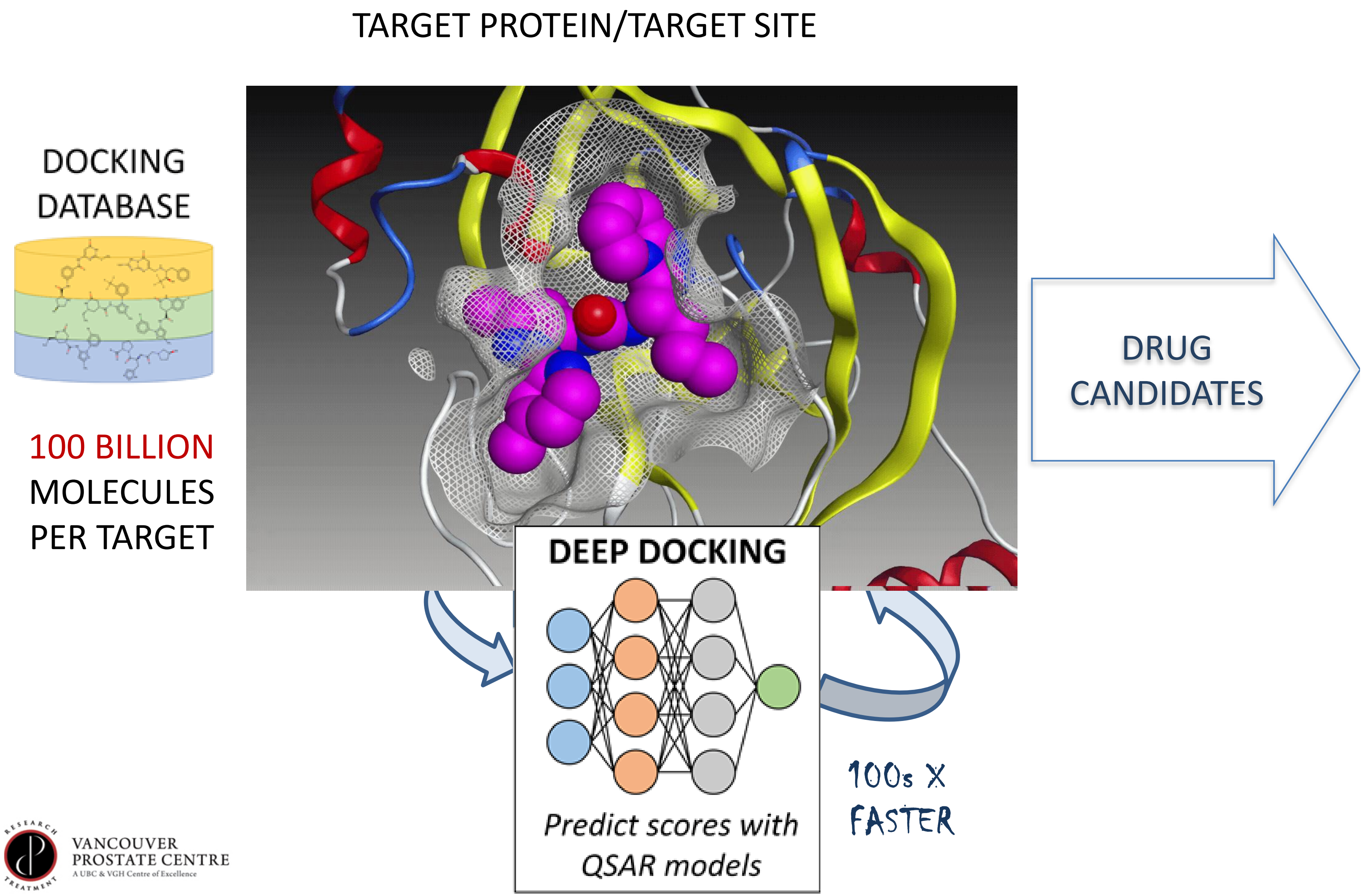


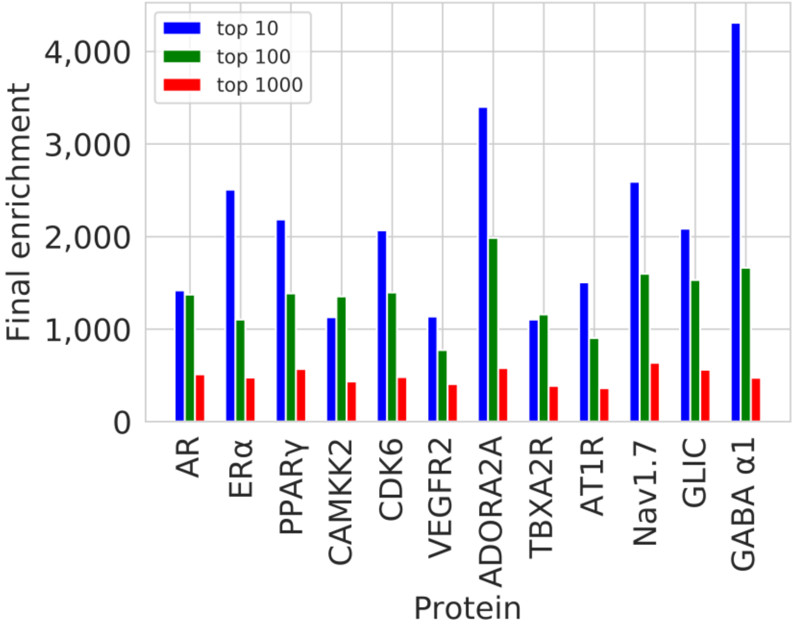
PREDICTED HIGH SCORING MOLECULES AUGMENT THE TRAINING SET OF THE MODEL (1% IN TOTAL)

ACTIVE/INACTIVE CUT-OFF TO IS MADE MORE STRINGENT AT EVERY ITERATION

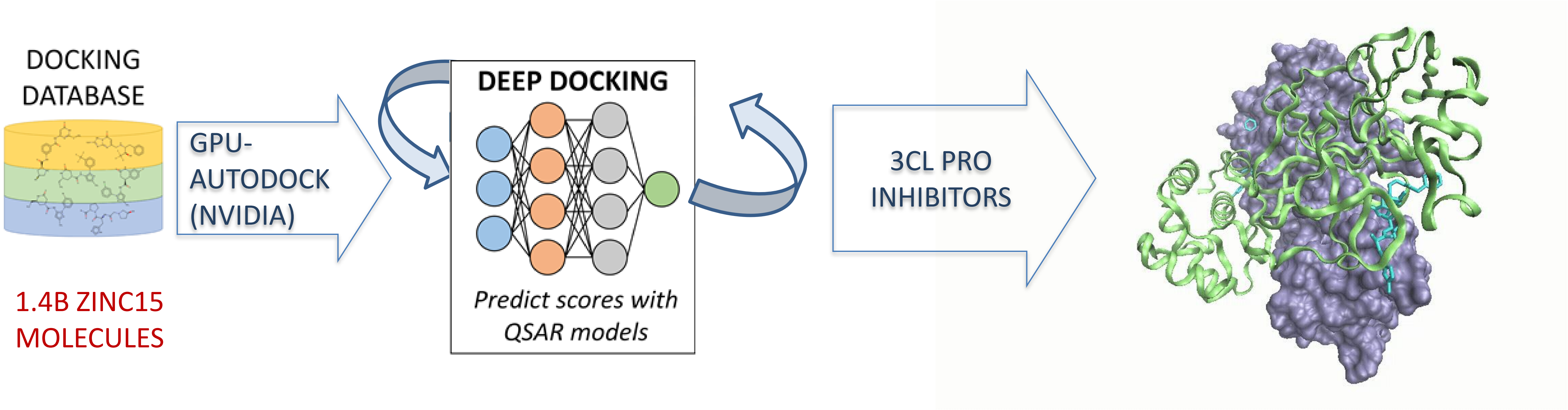
NR OF MOLECULES PREDICTED AS VIRTUAL HITS AFTER EACH ITERATION IS REDUCED

DEEP DOCKING PROVIDES 1000-S FOLD ACCELERATION OF VIRTUAL SCREENING


TARGET PROTEIN/TARGET SITE
DOCKING DATABASE
100 BILLION MOLECULES PER TARGET
DEEP DOCKING
Predict scores with QSAR models
100s X FASTER
DRUG CANDIDATES



DEEP DOCKING FOR SARS-COV-2 DRUG DISCOVERY



DOCKING DATABASE

GPU-AUTODOCK (NVIDIA)

DEEP DOCKING

Predict scores with QSAR models

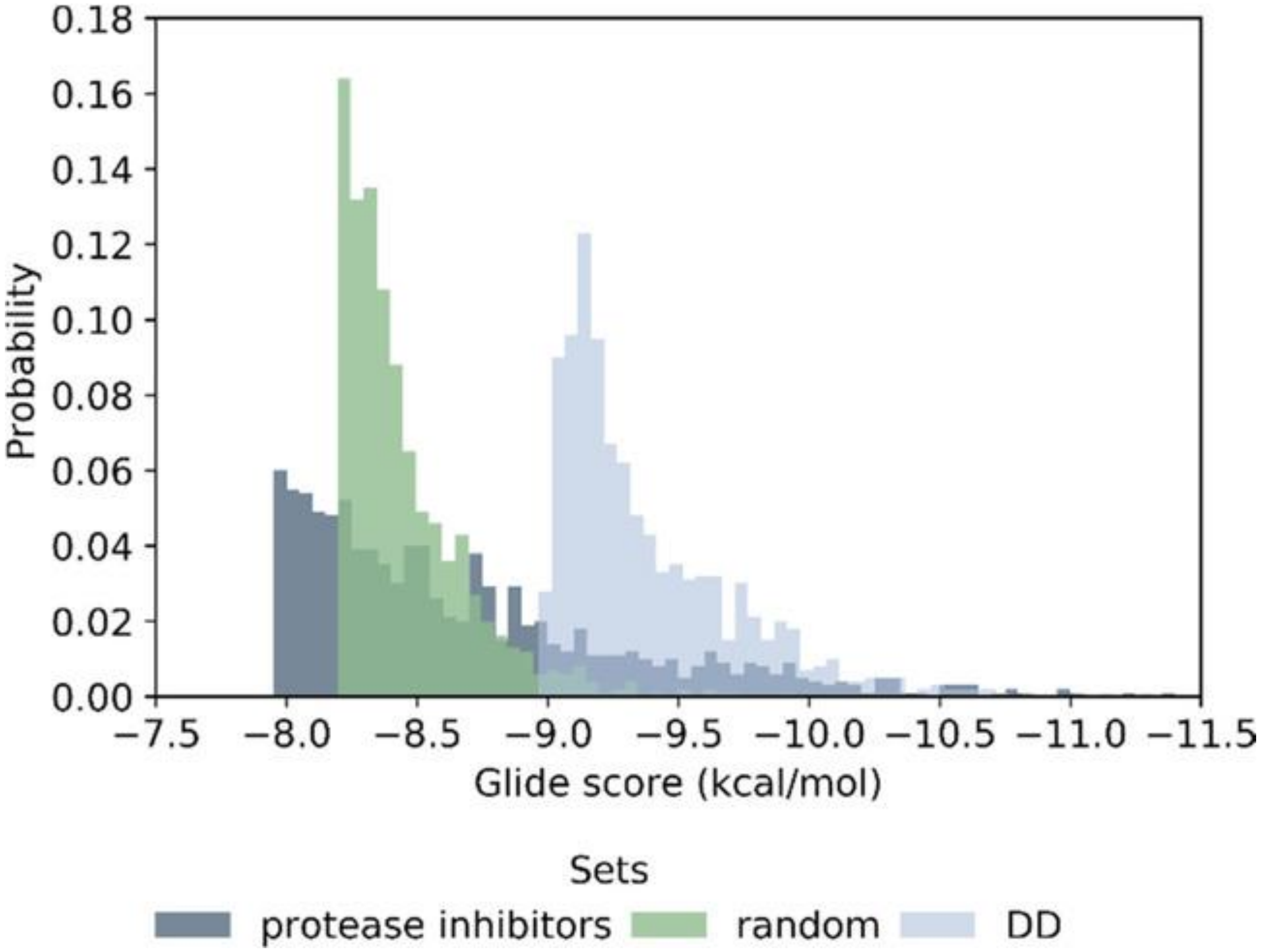
3CL PRO INHIBITORS

1.4B ZINC15 MOLECULES

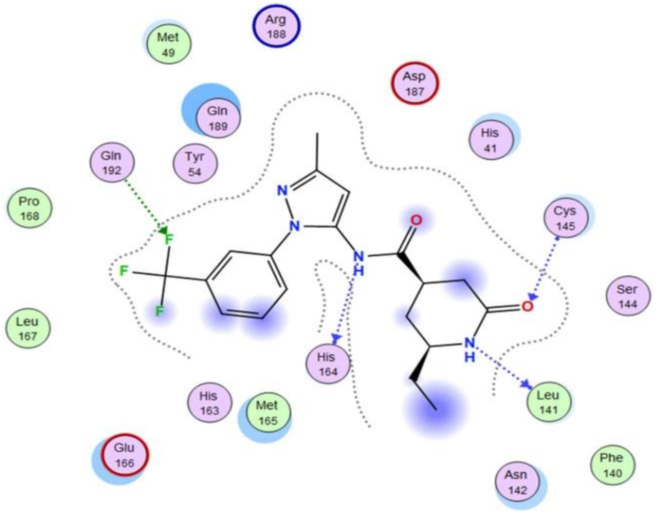
SARS-COV-2 3CL PROTEASE

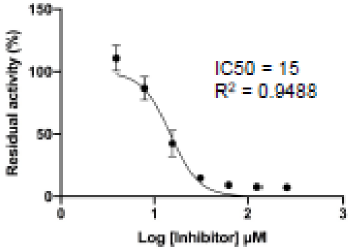
DEEP DOCKING IDENTIFIED 585 POTENTIAL 3CL PRO INHIBITORS

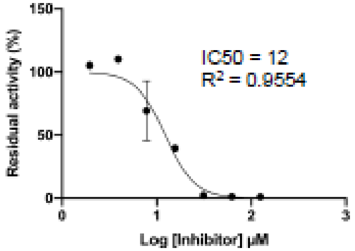
DOCKING SCORES OF TOP 1,000 CANDIDATES SIGNIFICANTLY BETTER THAN OF KNOWN BENCHMARKS



30+ INHIBITORS OF 3CL PRO ENZYME ARE CONFIRMED ACTIVE





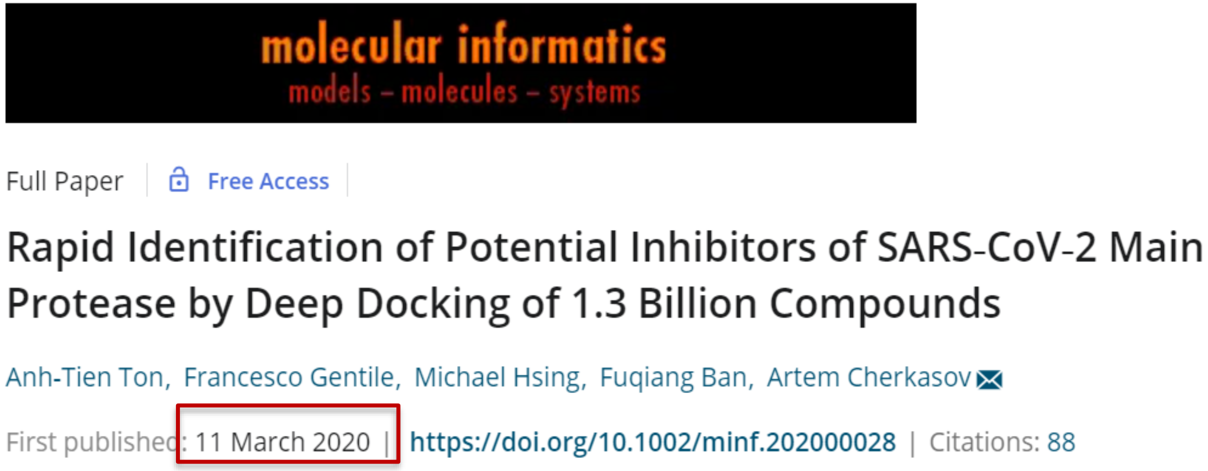


OUR FIRST PUBLICATION WITH INITIAL DRUG CANDIDATES

AGAINST COVID19 APPEARED AS EARLY AS **FEB19, 2020**

1,000 CANDIDATE 3CL PRO INHIBITORS

DISCLOSED TO THE PUBLIC



molecular informatics
models – molecules – systems

Full Paper | Free Access

Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds

Anh-Tien Ton, Francesco Gentile, Michael Hsing, Fuqiang Ban, Artem Cherkasov

First published: 11 March 2020 | https://doi.org/10.1002/minf.202000028 | Citations: 88

OUT OF 585 PREDICTED COMPOUNDS 30+ ACTIVE (5%)

BILLION-MOLECULES DRUG DISCOVERY

LARGER DOCKING LIBRARIES YIELD BETTER AND MORE HITS (LYU ET AL, NATURE, 2019)

MANY METHODS FOLLOWED OUR 2020 PAPER ON SCREENING 1B+ MOLECULES

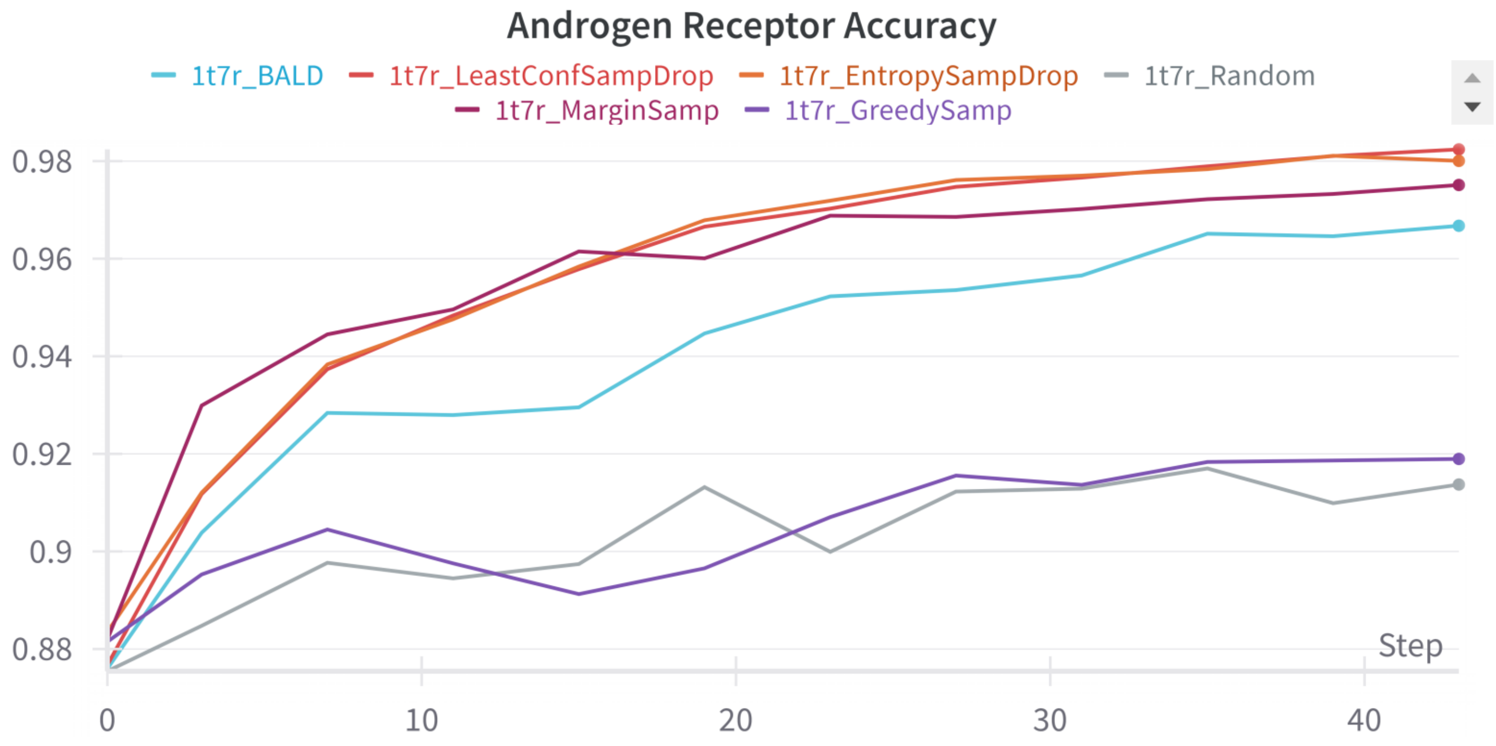
METHOD	REQUIRED TIME	SERVERS	DOCKING PROGRAM	TARGET	REFERENCE
OPENEYE ORION	<1 WEEK	45,000	FRED	PNP/HSP90	HTTPS://WWW.EYESOPEN.COM/ORION
AUTODOCK-GPU	<1 WEEK	27,600	AUTODOCK-GPU	SARS-COV-2 MPRO	ACHARYA ET AL, CHEMRXIV, 2020
VIRTUALFLOW	4 WEEKS	8,000	QUICKVINA, VINA, …	KEAP1-NRF2 INTERACTION	GORGULLA ET AL, NATURE, 2020
DEEP DOCKING	5 WEEKS	4	FRED, GLIDE	MULTIPLE TARGETS	GENTILE ET AL, CENTRAL SCIENCE, 2020

COMPARING ACQUISITION FUNCTIONS ON DOCKING DATA

Dataset: Random 3M ZINC compounds docked to ANDROGEN RECEPTOR LIGAND-BINDING DOMAIN (PDB: 1T7R)

Task:

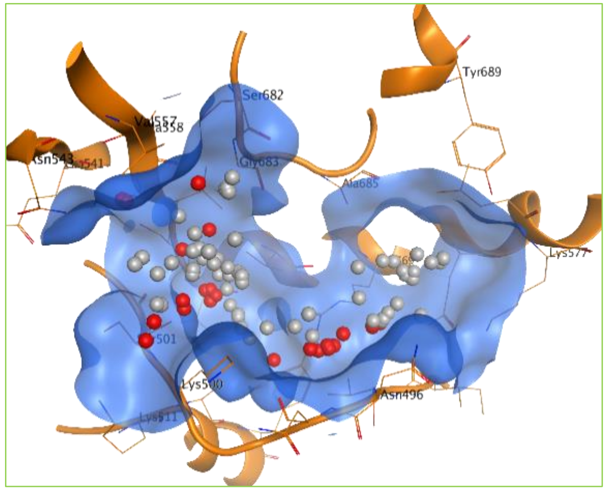
- Build a classification model to distinguish good binders from bad.
- Demonstrate the effectiveness of uncertainty based acquisition functions over greedy acquisition.



Uncertainty-based acquisition functions, such as MarginSampling, EntropySampling, and Bayesian Active Learning with dropout, improve model performance over the GreedySampling approach.

FULLY AUTOMATED DOCKING
WITHOUT "EXPERT IN THE LOOP"

SARS-COV-2 3CL PRO TARGET

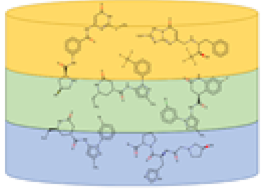


40B MOLECULES

ENAMINE R.S.

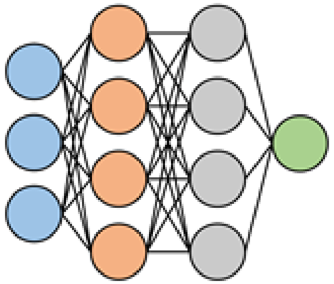
DOCKING

DATABASE



5 PROGRAMS

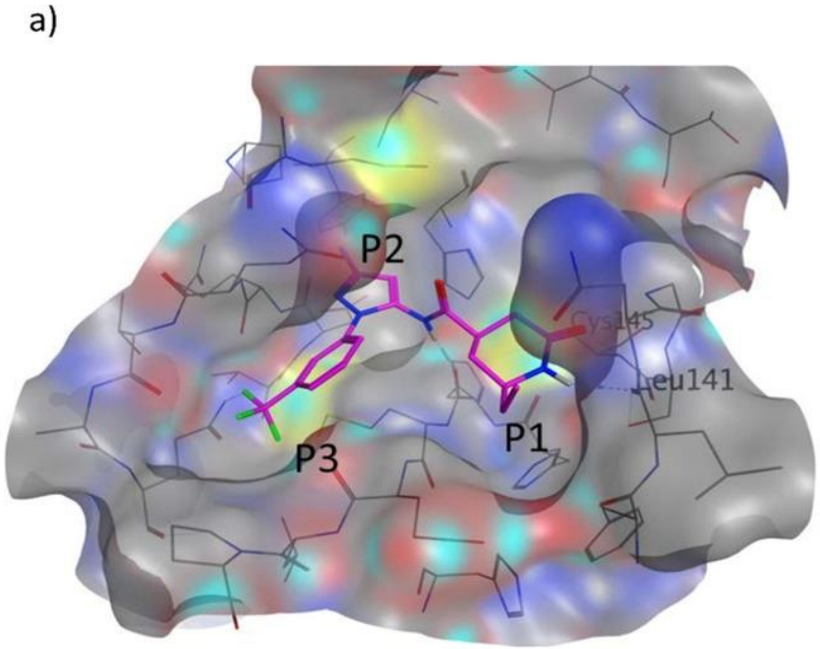
DEEP DOCKING



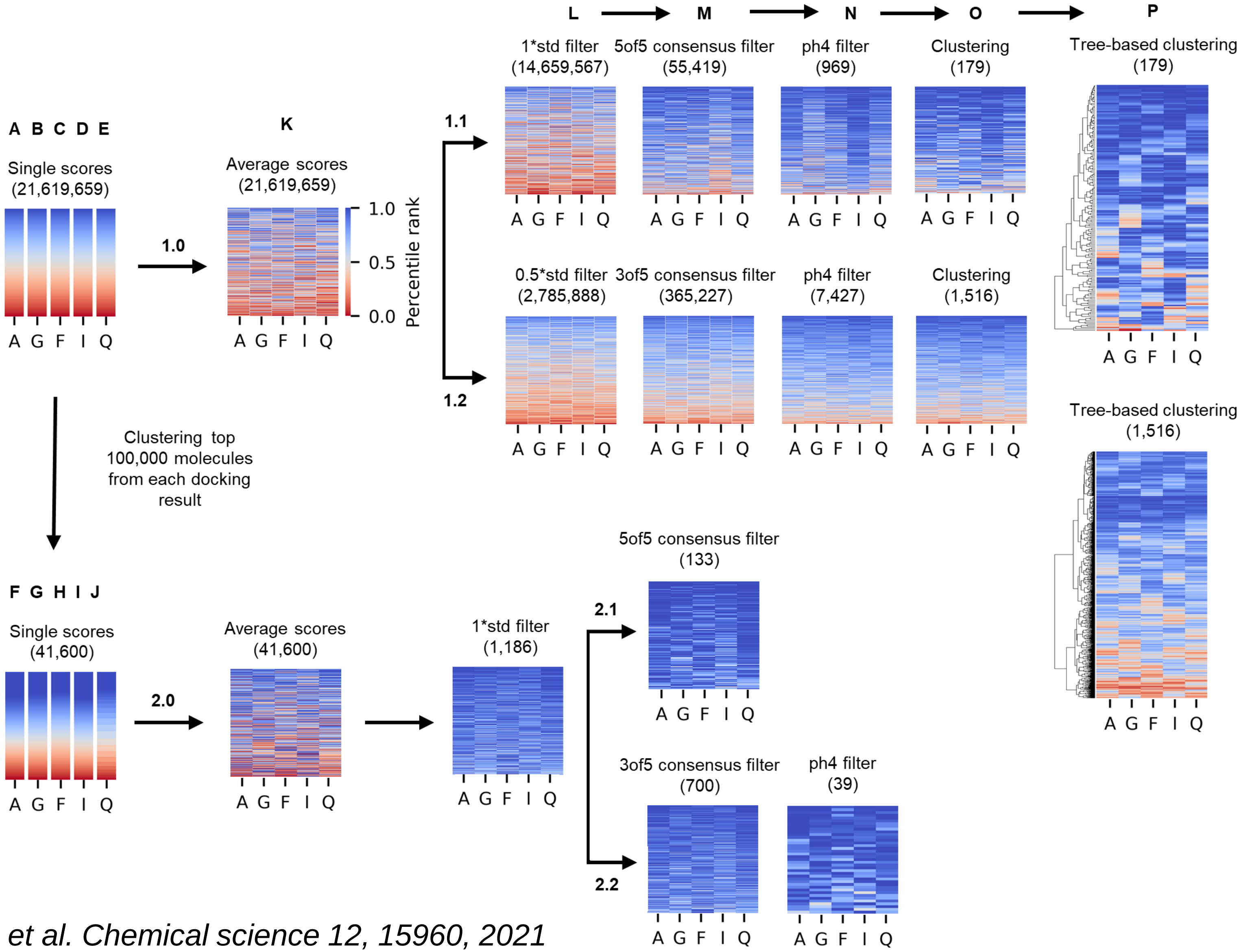
Predict scores with QSAR models

DRUG CANDIDATES

FROM 200 BILLION DOCKING RUNS

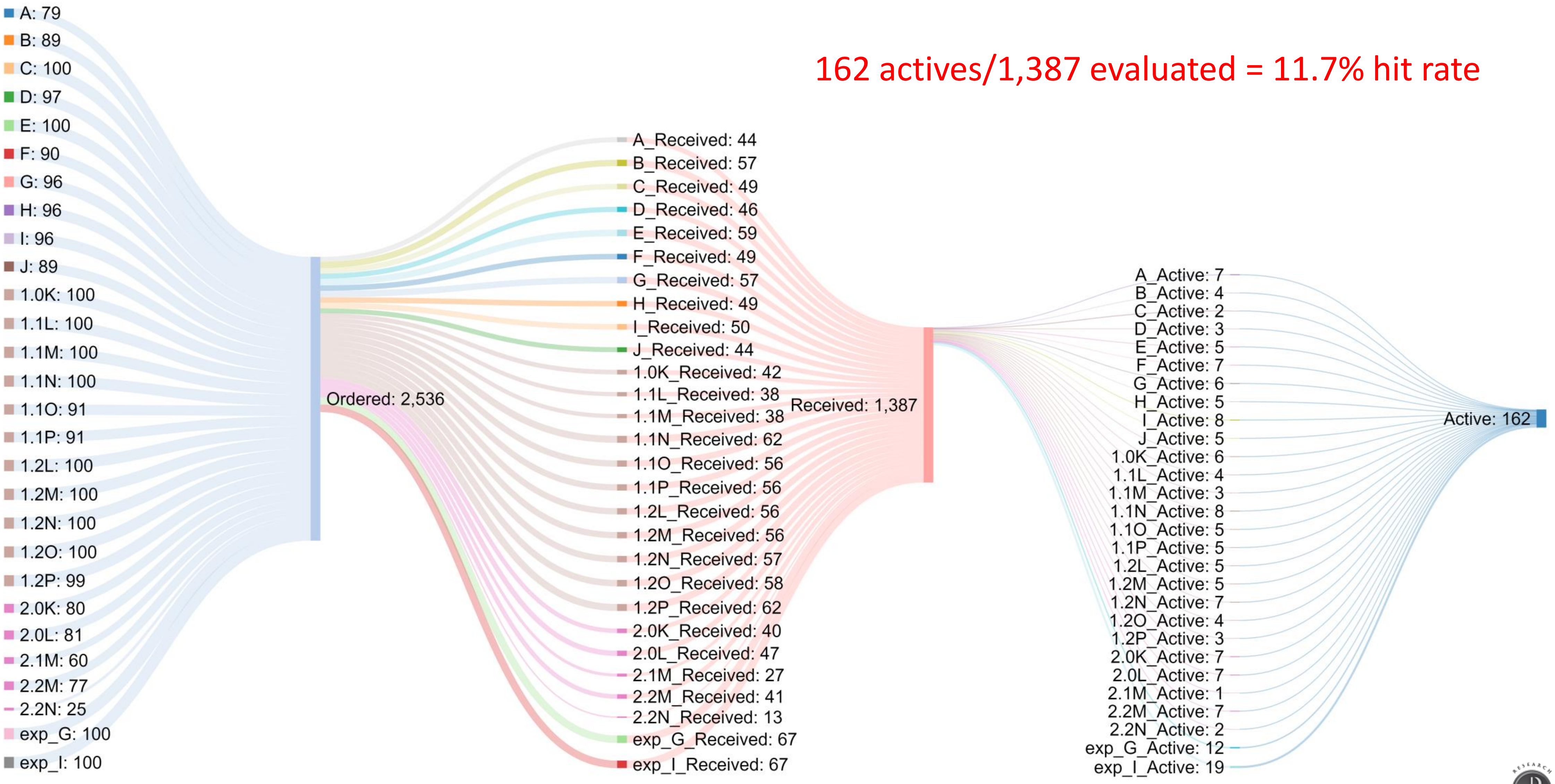


NO “EXPERT IN THE LOOP”



F. Gentle et al. Chemical science 12, 15960, 2021

AUTOMATED AND EXPERT-IN-THE-LOOP HIT RATES



162 actives/1,387 evaluated = 11.7% hit rate

F. Gentle et al. Chemical science 12, 15960, 2021

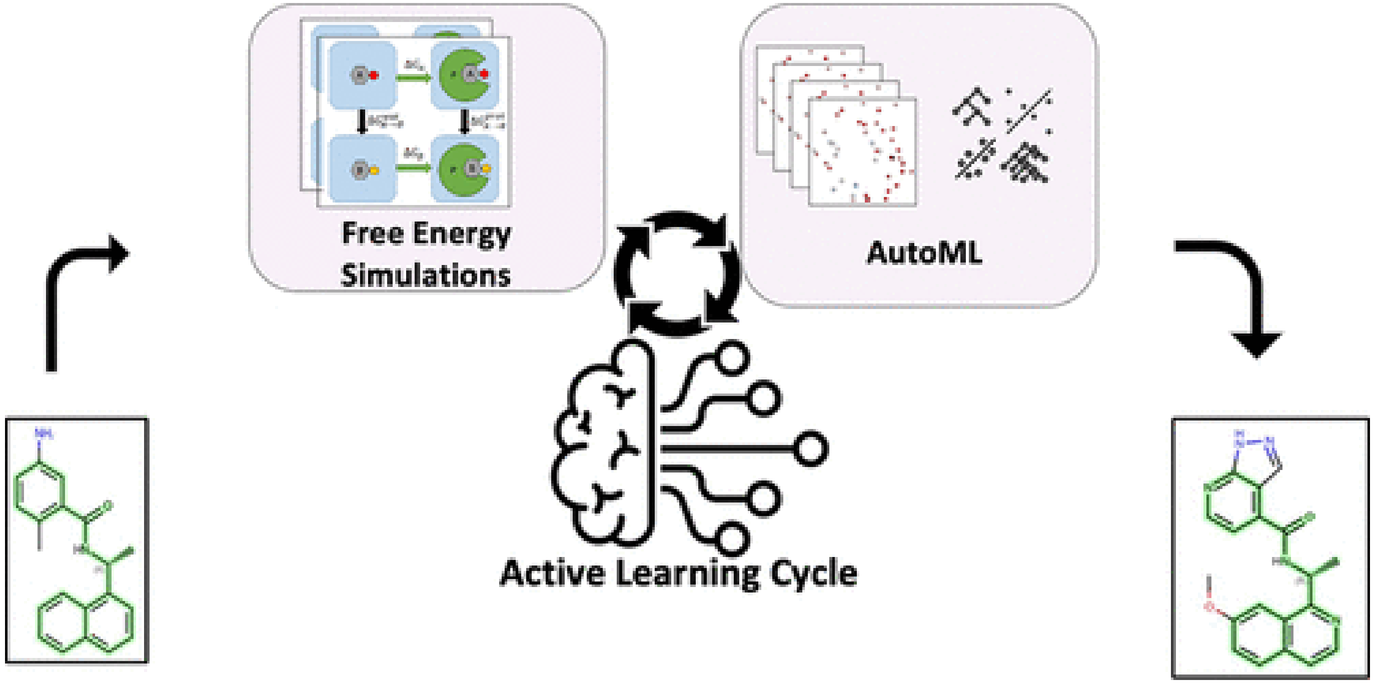
SIMILAR APPROACHES EMERGED

Method	Emulated docking score	Descriptors	QSAR function	References
DEEP DOCKING	GLIDE SP Quick Vina2 FRED GPU-AutoDock ICM	Morgan fingerprints	Deep Neural Network	105
Pyzer-Knapp approach	AutoDock-Vina	Extended connectivity fingerprints	Bayesian optimization	126
Jastrzebski et al approach	GLIDE XP SMINA	Contact fingerprints	Deep Neural Network	127
MolPal	AutoDock-Vina	Morgan fingerprints	Neural network Random forest Message passing neural network	128
MFP approach	DOCK	Morgan fingerprints	Linear regression	129
LEAN-DOCKING	GOLD AutoDock-Vina FRED GLIDE SP MOE	Unfolded counted atom pairs fingerprints	Regressor model	130
HASTEN	GLIDE SP FRED	Morgan fingerprints	Message passing neural network	131

MEMES	AutoDock	Extended connectivity fingerprints;
		Mol2Vec descriptors;
		CDDD descriptors
			Convolutional neural network
			Recurrent neural network
				132

Yang et al approach	GLIDE SP DOCK 3.7	Morgan fingerprints; Molecular graphs	Graph-Convolutional Neural Network Random forest	133
V-DOCK	AutoDock-Vina	2048 RDKit fingerprints combined with 166 bits MACSS fingerprints	PyTorch deep learning library	134
Bucinsky et al approach	AutoDock	SOAP molecular descriptors; SchNet 128 bits vectors	Keras neural network Deep tensor neural network Gradient boosted decision tree	135
NeuralDock	MedusaDock	36 bits atom type vectors with 7 channels for ligands; 10 × 10 × 10, 2-angstrom resolution images with 8 channels for protein pockets	TensorFlow Neural Network	136
MILCDOCK	LeDock PLANT Vina AutoDock 4 rDock	Pose-based RMSD values; Docking programs' metadata	Gradient boosted trees Random forest Naïve Bayes Neural Network	137
DOCKSTRING	AutoDock-Vina	Various fingerprints	Regressions Gradient boosted trees Gaussian processes Graph neural network	138
etc	etc	etc	etc	

AL-AUTOML WORKFLOW PROVEN IN DIFFERENT CONTEXT



Free Energy Simulations

AutoML

Active Learning Cycle

100x potency improvement with 3% of library sampled

January 23, 2023 Volume 63, Issue 2
pubs.acs.org/jcim
JCIM
JOURNAL OF CHEMICAL INFORMATION AND MODELING
ACS Publications
Most Trusted. Most Cited. Most Read.
www.acs.org

Filipp Gusev, Evgeny Gutkin, Maria G. Kurnikova, and Olexandr Isayev. **Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling** *J. Chem. Inf. Model.* 2023, 63, 2, 583−594. https://doi.org/10.1021/acs.jcim.2c01052

Slide courtesy CMU

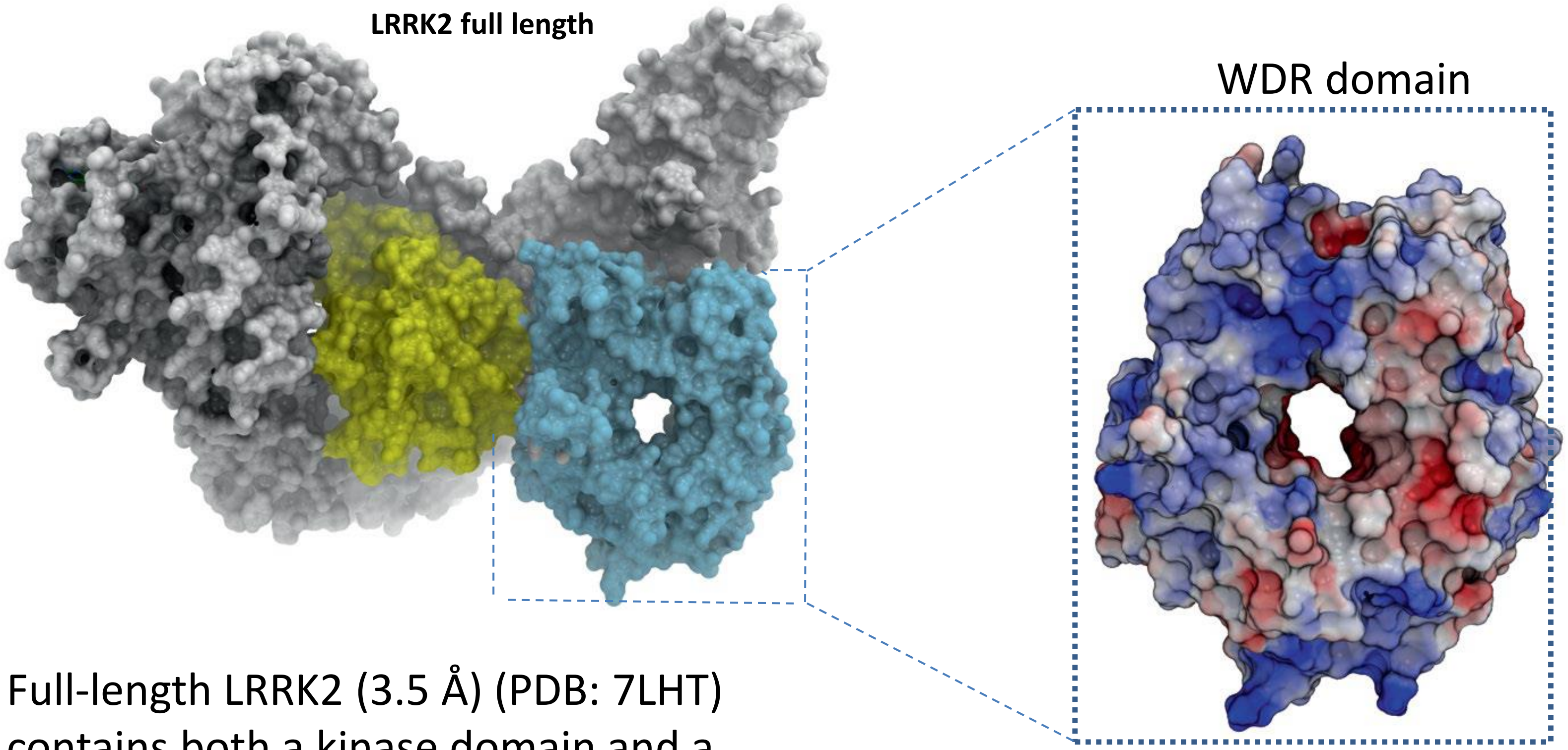
LATEST INITIATIVES

CACHE-1

- A public benchmarking project to compare and improve small-molecule hit-finding algorithms through cycles of prediction and experimental testing

- LRRK2 WDR: Potential Drug target for familial Parkinson's Disease

- No known small molecule inhibitors



Full-length LRRK2 (3.5 Å) (PDB: 7LHT) contains both a kinase domain and a WD40 repeat (WDR) domain.

LRRK2 WDR domain (2.7 Å) [PDB: 6DLO]

CACHE CHALLENGE

https://cache-challenge.org/

Slide courtesy CMU

CACHE-1 TEAM: ACTIVE LEARNING^2

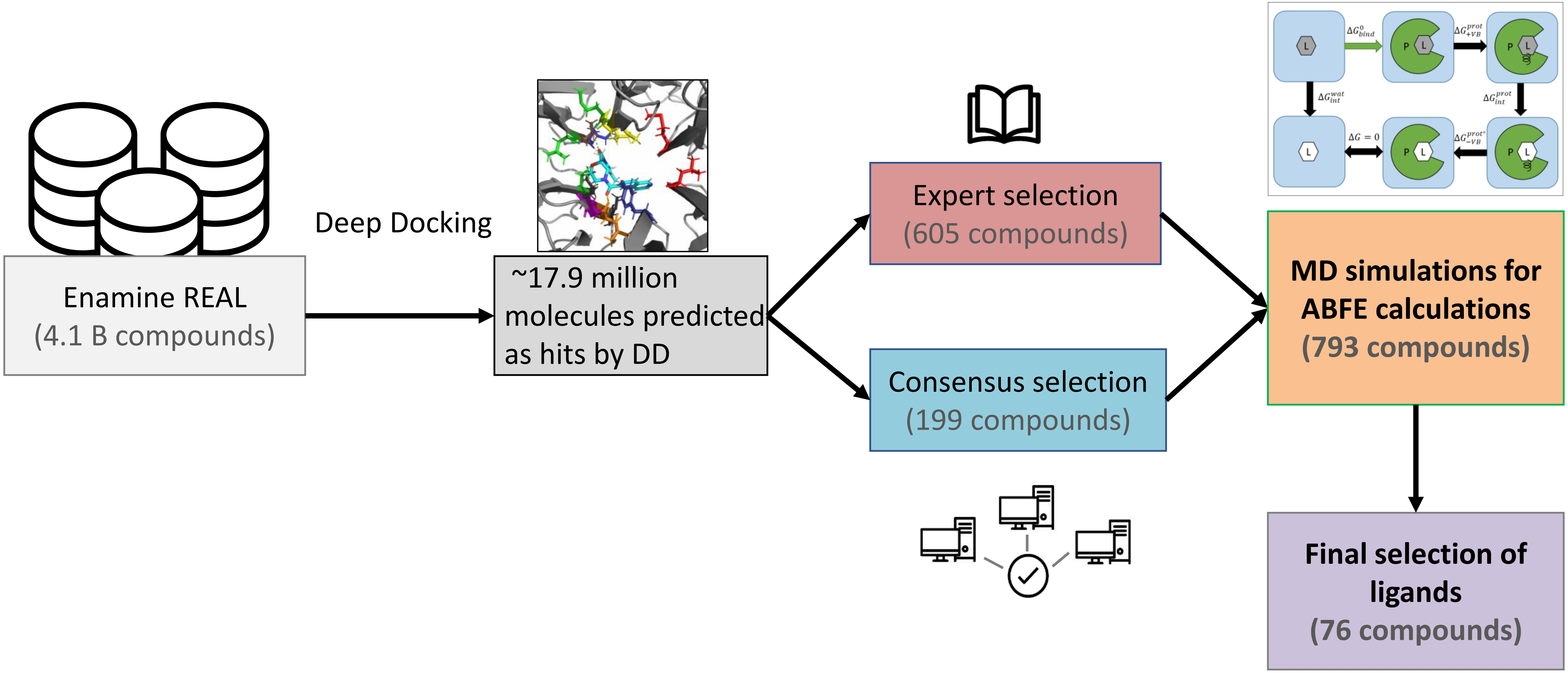
UBC

uOttawa

Carnegie Mellon University

Slide courtesy CMU

OVERVIEW OF THE ROUND 1 PIPELINE



Enamine REAL
(4.1 B compounds)

Deep Docking

~17.9 million molecules predicted as hits by DD

Expert selection
(605 compounds)

Consensus selection
(199 compounds)

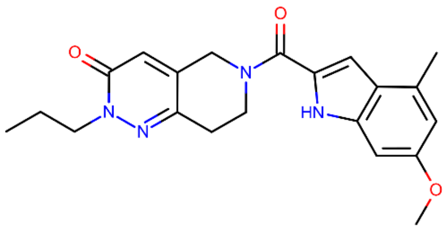
MD simulations for ABFE calculations
(793 compounds)

Final selection of ligands
(76 compounds)

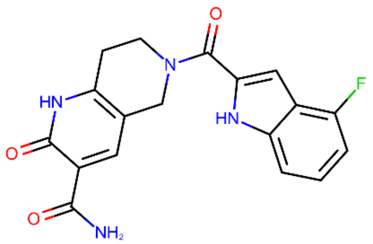
Gutkin E, Gusev F, Gentile F, Ban F, Koby SB, Narangoda C, *et al.* In silico screening of LRRK2 WDR domain inhibitors using deep docking and free energy simulations. *ChemRxiv*. **2024**; doi:10.26434/chemrxiv-2023-lnzvr

Slide courtesy CMU

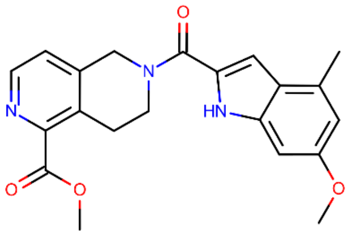
EXPERIMENTALLY VALIDATED HITS



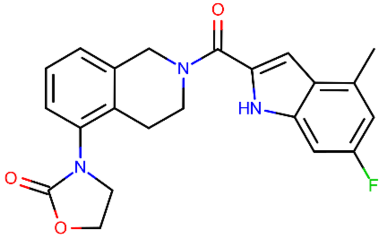
O1 K_d: 14.0μM



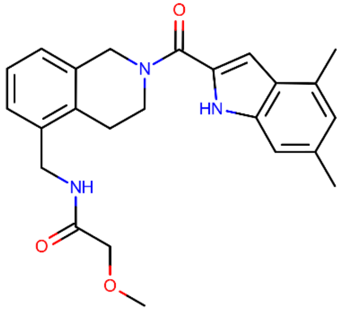
O2 K_d: 19.0μM



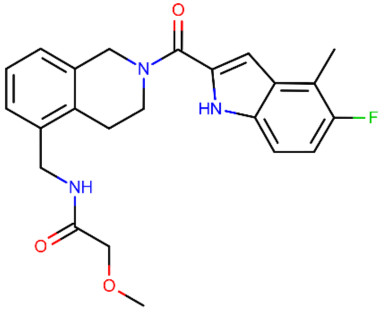
O3 K_d: 19.3μM



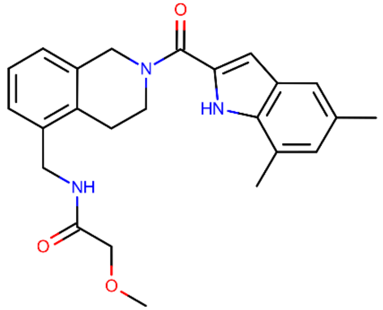
O4 K_d: 65.3μM



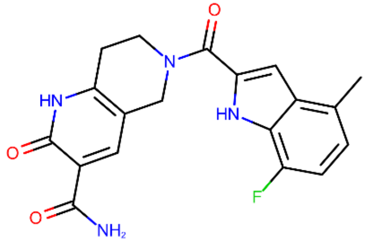
O5 K_d: 67.8μM



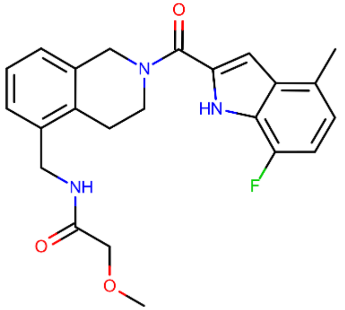
O6 K_d: 108.0μM



O7 K_d: 117.0μM



O8 K_d: 142.0μM



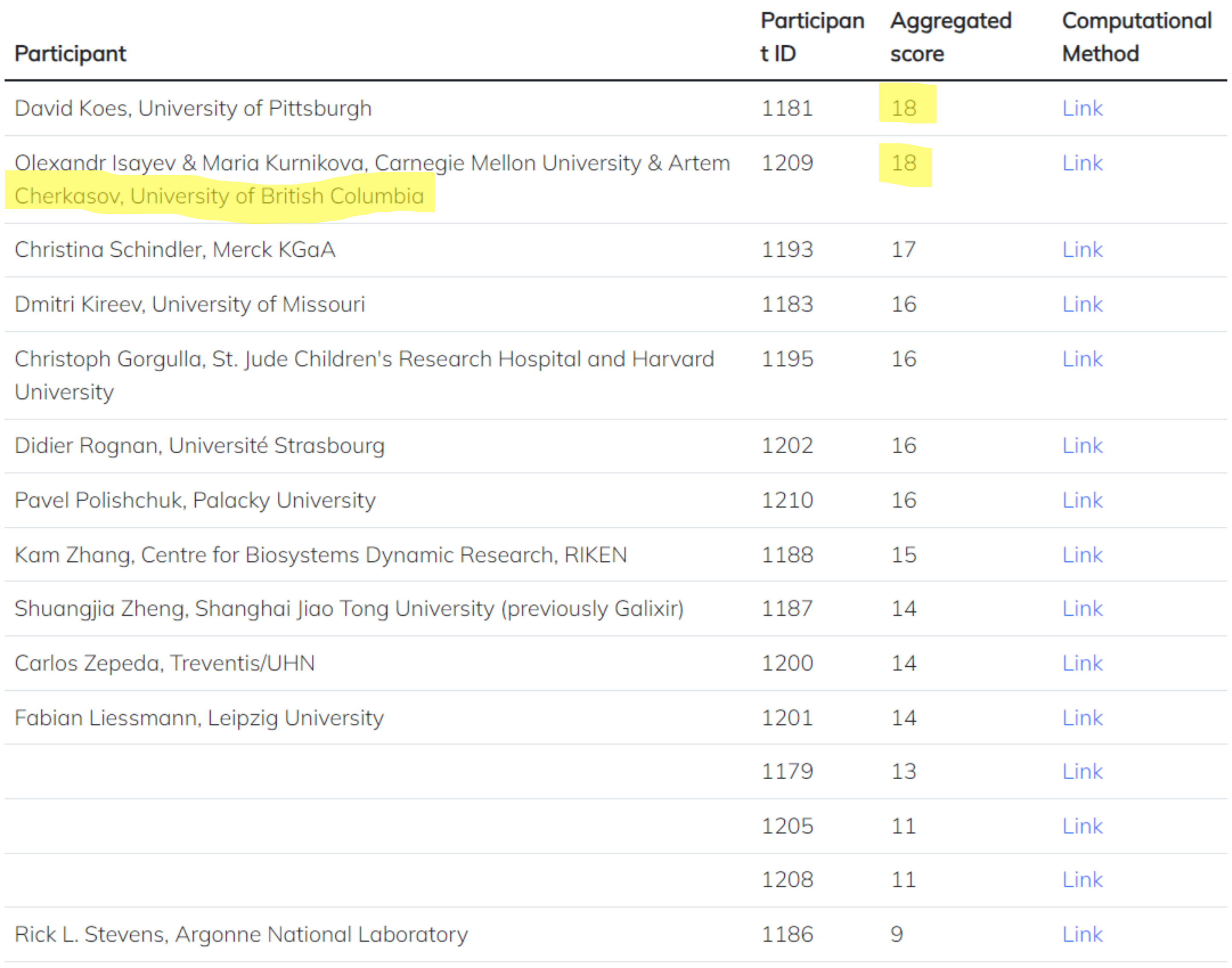
O9 K_d: 249.0μM

hit	ΔΔG	K_d (hit 1) / K_d
O1	-2.1	34.3
O2	-1.92	25.3
O3	-1.91	24.9
O4	-1.18	7.4
O5	-1.16	7.1
O6	-0.89	4.4
O7	-0.84	4.1
O8	-0.72	3.4
O9	-0.39	1.9

K_d (hit 1) = 480 μM

Slide courtesy CMU

CRITICAL ASSESSMENT OF COMPUTATIONAL HIT-FINDING EXPERIMENTS



Participant	Participant ID	Aggregated score	Computational Method
David Koes, University of Pittsburgh	1181	18	Link
Olexandr Isayev & Maria Kurnikova, Carnegie Mellon University & Artem Cherkasov, University of British Columbia	1209	18	Link
Christina Schindler, Merck KGaA	1193	17	Link
Dmitri Kireev, University of Missouri	1183	16	Link
Christoph Gorgulla, St. Jude Children's Research Hospital and Harvard University	1195	16	Link
Didier Rognan, Université Strasbourg	1202	16	Link
Pavel Polishchuk, Palacky University	1210	16	Link
Kam Zhang, Centre for Biosystems Dynamic Research, RIKEN	1188	15	Link
Shuangjia Zheng, Shanghai Jiao Tong University (previously Galixir)	1187	14	Link
Carlos Zepeda, Treventis/UHN	1200	14	Link
Fabian Liessmann, Leipzig University	1201	14	Link
	1179	13	Link
	1205	11	Link
	1208	11	Link
Rick L. Stevens, Argonne National Laboratory	1186	9	Link

23 finalists including

Merck

Bayer

Boehringer Ingelheim

Harvard

Argonne Lab

etc…

TEAMS AND FUNDS

DR. FRANCESCO GENTLE

DR. FUQIANG BAN

DR. MICHALE LLAMOSA

DR. MICHAEL HSING

DR. ERIC LEBLANC

DR. JAMES SMITH

DR. CARL PEREZ

DR. NADA LALLOUS

DR. ALIERZA KHAN

DR. ANH-TIEN TON

HAZEM MSLATI

JAMES GLEAVE

JEAN CHARLE YAACOUB

MOHIT PANDEY

MARIIA RADAEVA

OLIVIA GARLAND

JIAYING YU

JANE FOO

DR. OLES ISAEV

DR. MARIA KURNIKOVA

PHIL GUSEV

EVGENY GUDKIN

BEN KOBY

DR. GERALDINE GUERON

DR. MARTINA CRISPO

DR. AYELÉN TORO

Institut Pasteur de Montevideo

CONICET IQUIBICEN UBA

UBC

CANADA'S DIGITAL TECHNOLOGY SUPERCLUSTER

NCE RCE Networks of Centres of Excellence of Canada | Réseaux de centres d'excellence du Canada

NATIONAL CANCER INSTITUTE

RESEARCH TREATMENT VANCOUVER PROSTATE CENTRE A UBC & VGH Centre of Excellence

TELUS

NSERC CRSNG

Prostate Cancer Foundation Curing Together.

Vancouver CoastalHealth Research Institute Healthier lives through discovery

CIHR IRSC Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

THE SAFEWAY FOUNDATION

INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

Teck

DELL

NATIONAL INSTITUTES OF HEALTH

FUNDED BY MOVEMBER

Prostate Cancer Canada

NVIDIA.

DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA

compute canada

Canadian Cancer Society Research Institute