nature protocols

https://doi.org/10.1038/s41596-024-00976-2

Protocol

Check for updates

Standardized production of hPSC-derived cardiomyocyte aggregates in stirred spinner flasks

In the format provided by the authors and unedited

Supplementary Information

Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking

Francesco Gentile[1], Jean Charle Yaacoub[1,#], James Gleave[1,#], Michael Fernandez[1], Anh-Tien Ton[1], Fuqiang Ban[1], Abraham Stern[2], Artem Cherkasov[1,*]

[1]*Vancouver Prostate Centre, Department of Urologic Sciences, The University of British Columbia, Vancouver, BC, Canada*

[2]*NVIDIA Corporation, Santa Clara, CA, United States of America*

[#]*Equal contribution*

[*]*Corresponding author, email: acherkasov@prostatecentre.com*

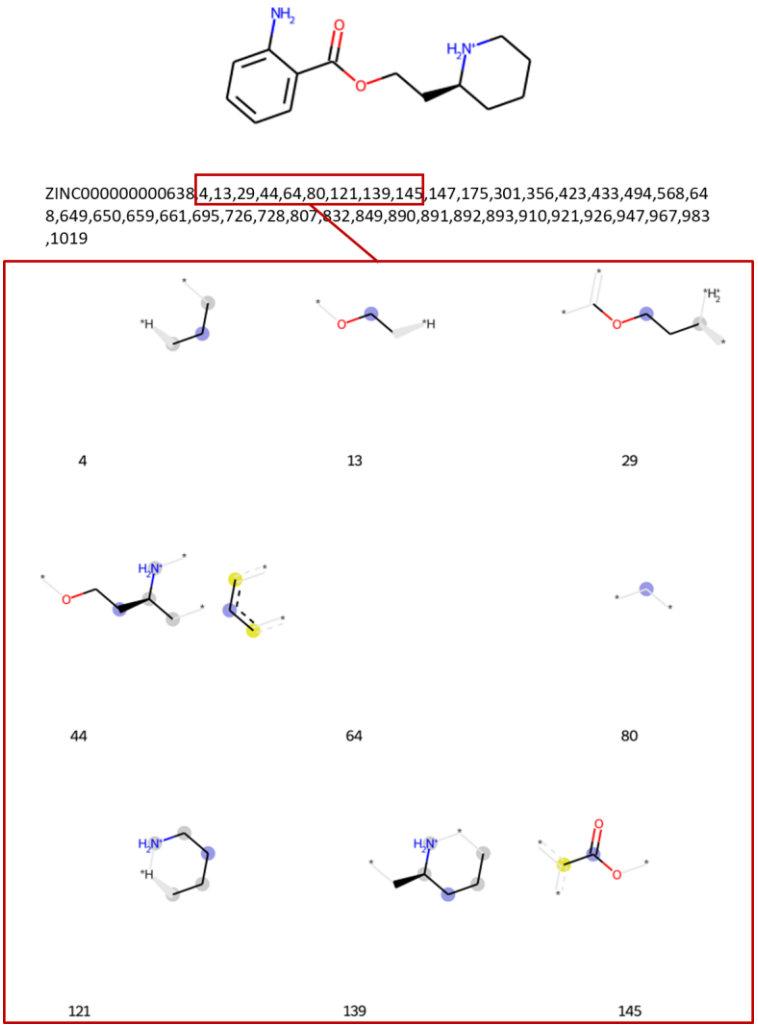
Supplementary Table 1. *Machine learning methods for accelerated docking-based virtual screening. SHBG, sex hormone-binding globulin; SARS 3CLPro, severe acute respiratory syndrome 3C-like protease; HIV, human immunodeficiency virus; LIT-PCBA, Laboratoire d'Innovation Thérapeutique - PubChem Assays data set; PTPN22, Protein tyrosine phosphatase, non-receptor type 22; MMP13, Collagenase 3 (Matrix Metallopeptidase 13); CTDSP1, Carboxy-terminal domain RNA polymerase II polypeptide A small phosphatase 1; TMPK, Thymidylate kinase; GPCR, G protein-coupled receptor; CYP, cytochrome P450; ADORA2A, G protein-coupled receptors included adenosine A2A receptor; AR, androgen receptor; AT1R, angiotensin II receptor type 1; CAMKK2, calcium/calmodulin-dependent protein kinase kinase 2; CDK6, cyclin-dependent kinase 6; ERα, estrogen receptor-alpha; GABAA, gamma-aminobutyric acid receptor type A; GLIC, Gloeobacter ligand-gated ion channel; Nav1.7, Nav1.7 sodium channel; PPARγ, peroxisome proliferator-activated receptor γ; TBXA2R, thromboxane A2 receptor; VEGFR2, vascular endothelial growth factor receptor 2; SARS-CoV-2 Mpro, severe acute respiratory syndrome coronavirus 2 main protease.*

Name	**Model**	**Descriptors**	**Targets**	**Software**	**Workload reduction**	**Size of library (1,000s)**	**Training set size**
Progressive Docking[1]	Partial least squares regression	Non-cross-correlating inductive and conventional QSAR descriptors	SHBG, SARS 3CLPro, HIV1 RT	Glide	1.2 - 2.6	90	10%
Lean-Docking[2]	Linear support vector regression	Signature molecular descriptors (2D graph based on atom's local environment in terms of neighboring atoms)[3]	LIT-PCBA data set	Gold, Vina, FRED, MOE-Dock, Glide	4 – 41	270	10,000
Spark CPVS[4]	Support vector machine and inductive conformal prediction	Signature molecular descriptors	HIV-1 protease, PTPN22, MMP13, CTDSP1	OEDocking TK	3.7	2,200	10%
MolPAL[5]	Directed-message passing neural network with Bayesian optimization	2,048-bit atom-pair fingerprints	TMPK	Vina	40	99,500	1%
Docking emulation[6]	Graph convolution network and edge-attention graph	Interaction fingerprint string descriptors	GPCRs and CYP enzymes	Glide	10	500	Depends on the number of active ligands

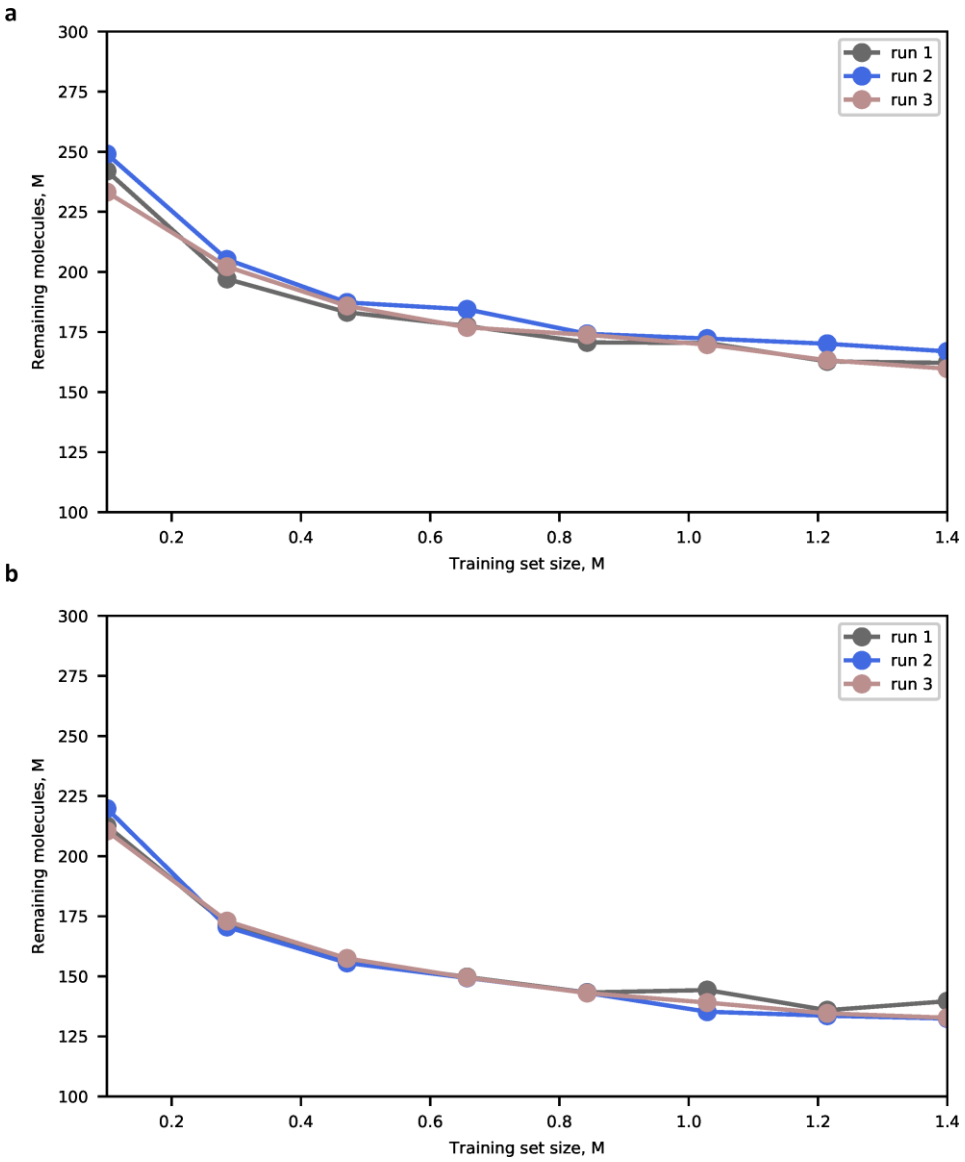
	convolution network						
Deep Docking (this protocol)[7]	Feed-forward deep neural network	1,024-bit Morgan fingerprints	ADORA, AR, AT1R, CAMK, CDK6, ERα, GABAA, GLIC, Nav1.7, PPARγ, TBXA2, VEGFR, SARS-CoV-2 Mpro	FRED, Glide	50-100	1,360,000	1%

Supplementary Table 2. *List of steps of Procedure 1 and 2.*

Process	**Steps in Procedure 1**	**Steps in Procedure 2**
Library preprocessing	1-3	1: perform steps 1-3 of Procedure 1
Library enumeration	4-6	2
Calculation of fingerprints	7	3
Receptor preparation	8-12	4: perform steps 8-12 of Procedure 1
Random sampling (DD phase 1)	13-18	5-7
Ligand preparation (DD phase 2)	19	8
Docking (DD phase 3)	20	9
Model training (DD phase 4)	21-23	10
Inference (DD phase 5)	24-31	11-12
Successive iterations	32-36	13-15: for step 14, perform step 35 of Procedure 1
Final phase	37-38	16-17: for step 17, perform step 38 of Procedure 1



Supplementary Figure 1. *Chemical structure of piridocaine (ZINC000000000638) (top) with its Morgan fingerprint in DD format. Framed in red: bit rendering (indicating presence of molecular substructures) for the first nine bits set to 1 for piridocaine. Central atoms are highlighted in blue, aromatic atoms in yellow, aliphatic atoms in dark gray. Fingerprint rendering was generated with rdkit (http://rdkit.blogspot.com/2018/10/using-new-fingerprint-bit-rendering-code.html).*



Supplementary Figure 2. *Comparison of the effect of training set size on the database reduction power of DD for docking simulations based on stochastic algorithms. The same molecular sets (training set of 1.4 million molecules, validation and testing sets of 700,000 molecules each) were docked to the active site of SARS-CoV-2 Mpro (PDB id 6W63[8]) using the Lamarckian genetic algorithm of Autodock-GPU[9] by setting the energy evaluation parameter (controlling the convergence of the docking) to* ***a****, 50,000 and* ***b****, 500,000 energy evaluations. For each setup, docking was performed three independent times. The obtained docking scores were used to train models with increasing sizes of the training set. More 'deterministic' docking runs (using 500,000 energy evaluations) led to a significantly better database reduction power at every tested sample size, compared to the 'random' runs (using 50,000 energy evaluations).*

Supplementary References

1. Cherkasov, A., Ban, F., Li, Y., Fallahi, M. & Hammond, G. L. Progressive docking: A hybrid QSAR/docking approach for accelerating in silico high throughput screening. *J. Med. Chem.* **49**, 7466–7478 (2006).

2. Berenger, F., Kumar, A., Zhang, K. Y. J. & Yamanishi, Y. Lean-Docking: Exploiting Ligands' Predicted Docking Scores to Accelerate Molecular Docking. *J. Chem. Inf. Model.* **61**, 2341–2352 (2021).

3. Faulon, J. L., Visco, D. P. & Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **43**, 707–720 (2003).

4. Ahmed, L. *et al.* Efficient iterative virtual screening with Apache Spark and conformal prediction. *J. Cheminform.* **10**, 8 (2018).

5. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).

6. Jastrzębski, S. *et al.* Emulating Docking Results Using a Deep Neural Network: A New Perspective for Virtual Screening. *J. Chem. Inf. Model.* **60**, 4246–4262 (2020).

7. Gentile, F. *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).

8. RCSB PDB - 6W63: Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77. https://www.rcsb.org/structure/6w63.

9. Santos-Martins, D. *et al.* Accelerating AutoDock4 with GPUs and Gradient-Based Local Search. *J. Chem. Theory Comput.* **17**, 1060–1073 (2021).