CompBioMed
A Centre of Excellence in Computational Biomedicine

Webinar series

Webinar #11
Next Generation Ultra High-Throughput Protein-Ligand Docking with Deep Learning

31 March 2020

The webinar will start at 5pm CEST

Presenter:
Austin Clyde (Argonne National Laboratory)

Moderator:
Apostolos Evangelopoulos (UCL)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675451

The webinar series is run in collaboration with:

VPH Institute
Building the Virtual Physiological Human

Webinar #11
Next Generation Ultra High-Throughput Protein-Ligand Docking with Deep Learning

31 March 2020

Welcome!

Presenter:
Austin Clyde (Argonne National Laboratory)

Moderator:
Apostolos Evangelopoulos (UCL)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675451

The webinar series is run in collaboration with:
VPH Institute
Building the Virtual Physiological Human

Next Generation Ultra High-Throughput Protein-Ligand Docking with Deep Learning

Utilizing a VAE to cluster states various ligands induce on a protein pocket [1]

AUSTIN CLYDE

Ph.D. Student, University of Chicago
Computational science, Argonne National Laboratory

aclyde@anl.gov
www.cs.uchicago.edu/~aclyde

Janssen
December 5, 2019

Acknowledgments

- This work was performed in part under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725

Overview

How can we screen 10 billion compounds on various protein targets in a reasonable, painless, and not expensive way?

1. Drug Discovery Pipeline
2. Giga-Docking
3. ML for drug discovery
4. Combining ML and Giga-Docking

Figure 1 | **Virtual screening today?** M. C. Escher's artwork "Ascending and Descending" from 1960 may be used to illustrate the current situation of virtual screening. With few exceptions (the potentially frustrated figures resting on the steps and on the balcony), the users and developers work on a high but similar level with sophisticated software. Although continuous step-by-step progress is made (or perceived as such), only a change of perspective will enable a breakthrough and allow computational chemistry to reach greater potential. Figure reproduced, with permission, from The M. C. Escher Company © (1960).

Overview of early stage drug discovery

DRUG DISCOVERY

~10^8 products

PRE CLINICAL

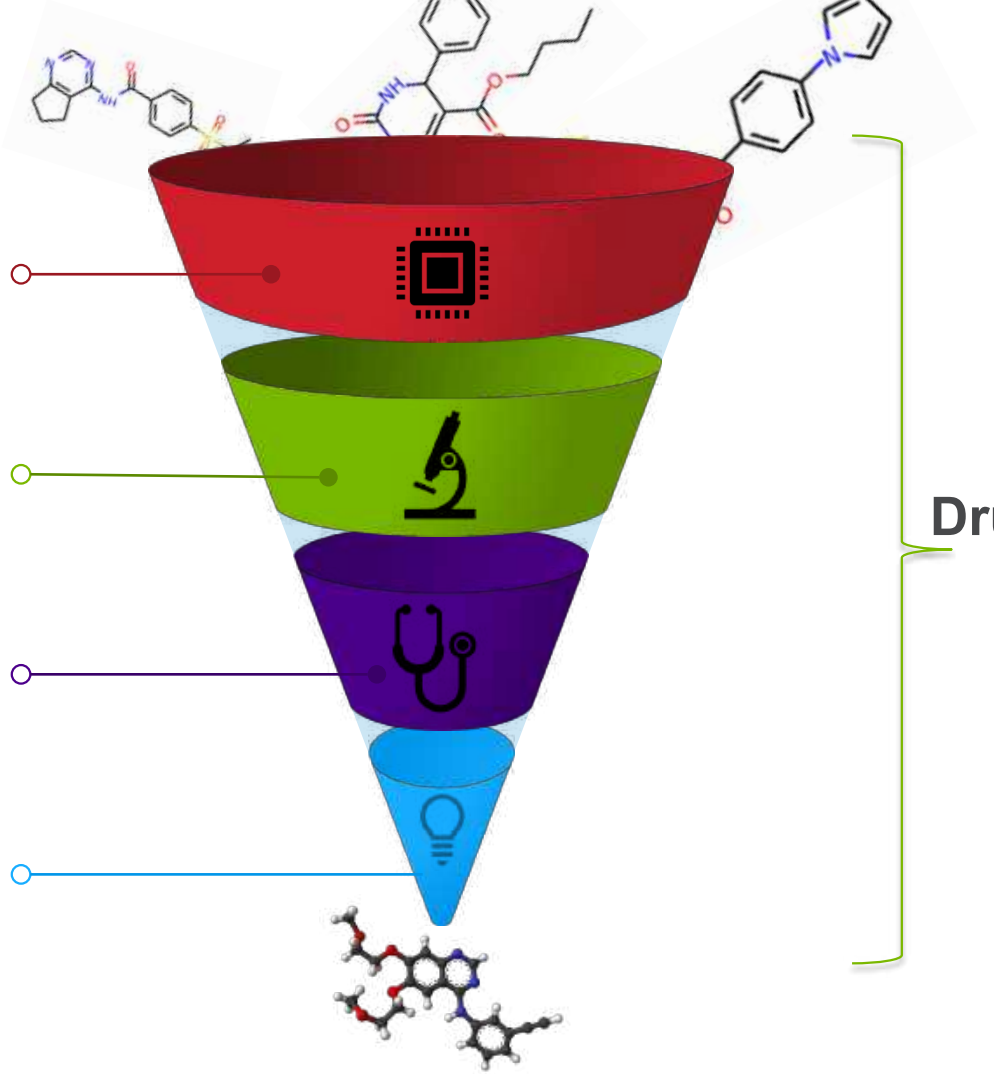
11,000 products

CLINICAL TRIALS

6,300 products

FDA APPROVAL

111 products



Drug Discovery Funnel

Target based compound screening

10^{60} estimated drug-like compounds

COMPOUND DISCOVERY

Mining massive building block or de-novo generated libraries

INTERESTING?

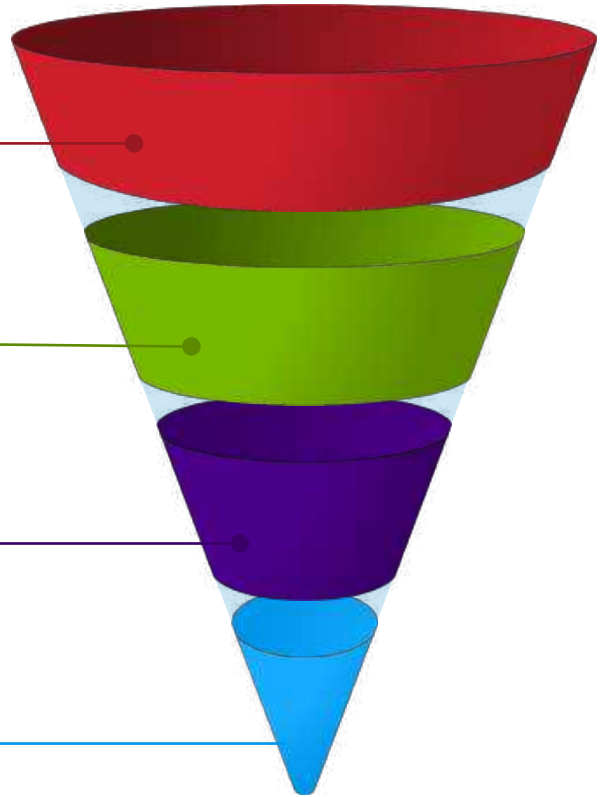
Does this compound inhibit or interact with the target?

TOXICOLOGY

Is this compound reasonably safe?

SYNTHESIS

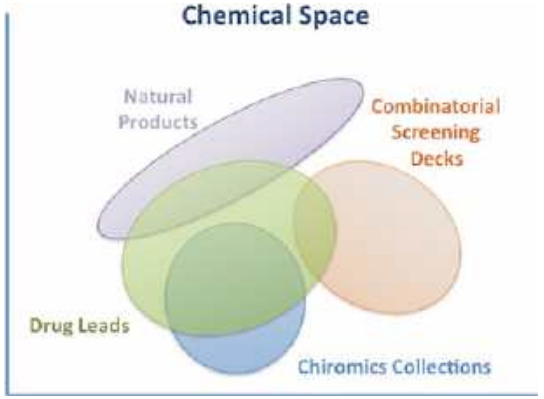
Can we buy it, is it from available building blocks, or do we need to hire a medicinal chemist?



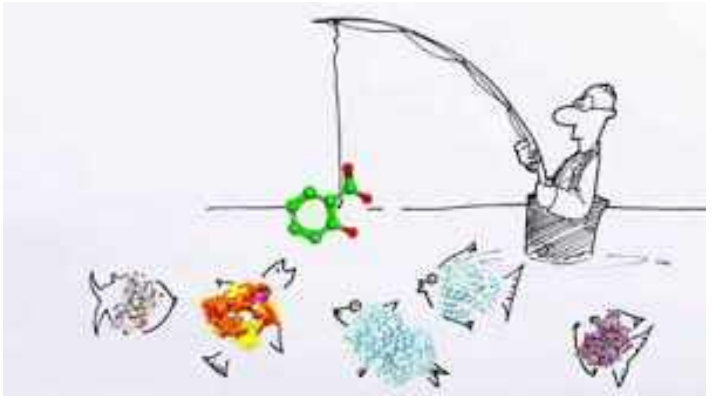
Lead Drug Discovery Funnel

Drug Discovery Pipeline

- Paul Ehrlich (1854-1915)
  - Magic Bullet: Ehrlich formed an idea that it could be possible to kill specific microbes which cause diseases in the body, without harming the body itself
- Estimated 2M proteins in the human body
- Estimated 10^{60} compounds

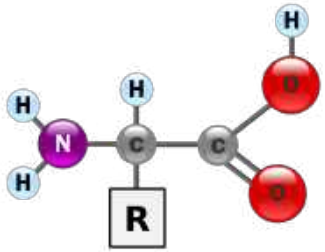


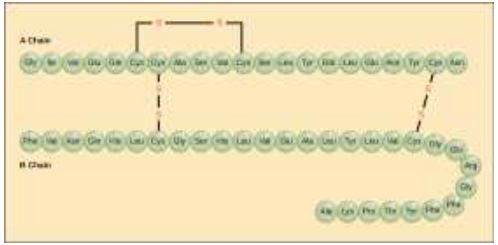
Searching for new drugs is like fishing in the dark: the prospect of catching something is very uncertain, and it requires patience, skill and - of course - money.

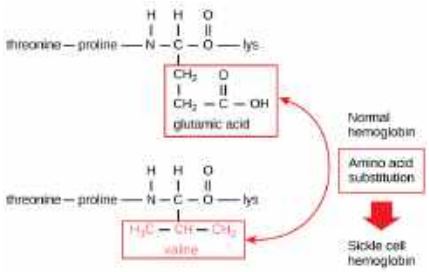


Video series from Roche youtu.be/bIFnOVKd2Ko

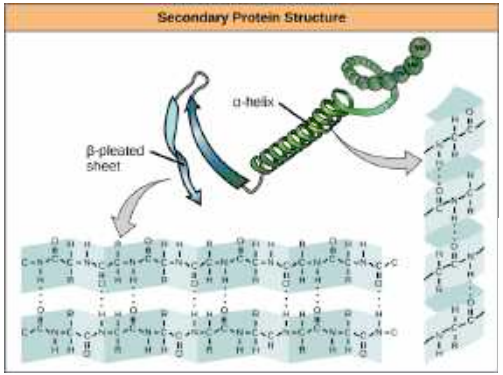
Proteins 101

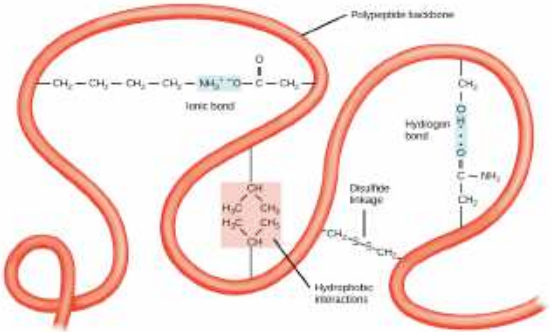






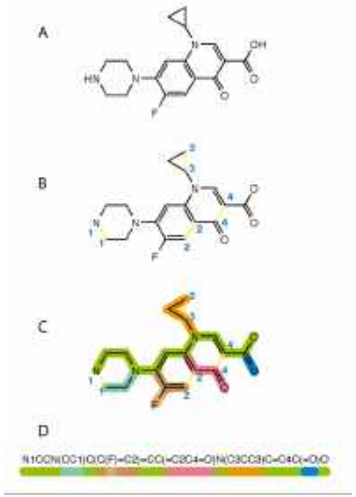
- **Primary structure**
  - sequence of amino acids in a polypeptide chain
- **Secondary** structure
  - refers to local folded structures that form within a polypeptide due to interactions between atoms of the backbone.
- **Tertiary** structure
  - The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.
- **Quaternary** structure
  - Multiple protein subunits come together to form a larger complex





Computational Chemistry

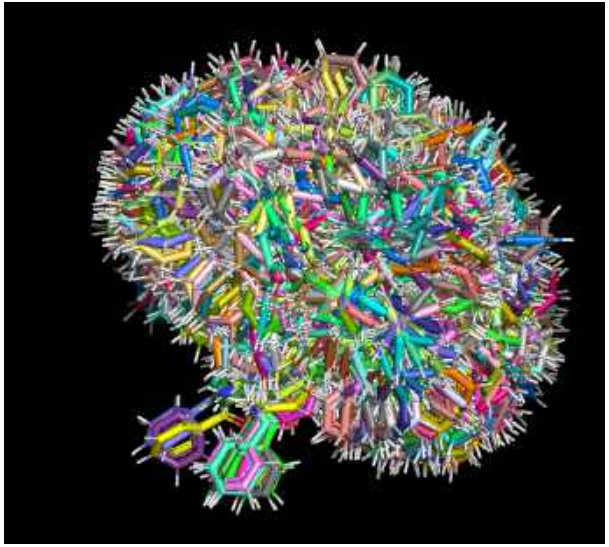
SMILES Simplified Molecular Input Line Entry System

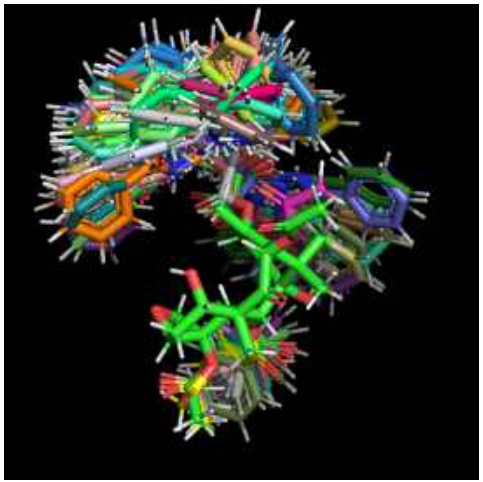


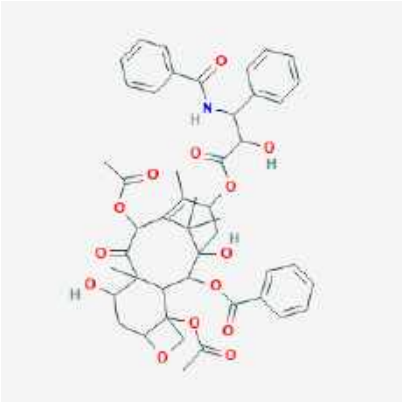
- De-facto standard for communicating molecular structures
  - It's somewhat difficult as different orders of the string exist
  - Streo0chemistry is hard
- DFS on graph
  - The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree.
  - Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

SMILES	Name	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine
c1ccccc1	benzene	N[C@H](C)C(=O)O	D-alanine

Input SMILES	Unique SMILES
OCC	CCO
[CH3][CH2][OH]	CCO
C-C-O	CCO
C(O)C	CCO
OC(=O)C(Br)(Cl)N	NC(Cl)(Br)C(=O)O
ClC(Br)(N)C(=O)O	NC(Cl)(Br)C(=O)O
O=C(O)C(N)(Br)Cl	NC(Cl)(Br)C(=O)O







"This evaluation has shown that docking programs are usually successful…the difficulty was not in positioning the ligand within the binding site but in reproduction of the small-molecule conformation."

A Critical Assessment of Docking Programs and Scoring Functions

Gregory L. Warren,*,† C. Webster Andrews,‡ Anna-Maria Capelli,# Brian Clarke,|| Judith LaLonde,†,§ Millard H. Lambert,‡ Mika Lindvall,⊥,● Neysa Nevins,† Simon F. Semus,† Stefan Senger,⊥ Giovanna Tedesco,# Ian D. Wall,|| James M. Woolven,⊥ Catherine E. Peishoff,† and Martha S. Head†

GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina 27709, GlaxoSmithKline, Centre Via Alessandro, FFleming 4, 37135, Verona, Italy, GlaxoSmithKline, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, U.K., and GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K.

Received April 17, 2005

Structural Docking

Exhaustive shape fitting

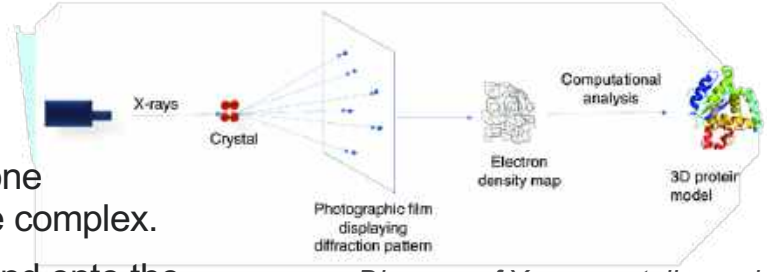
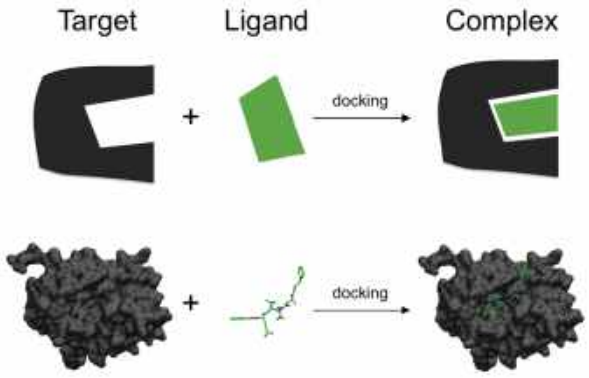


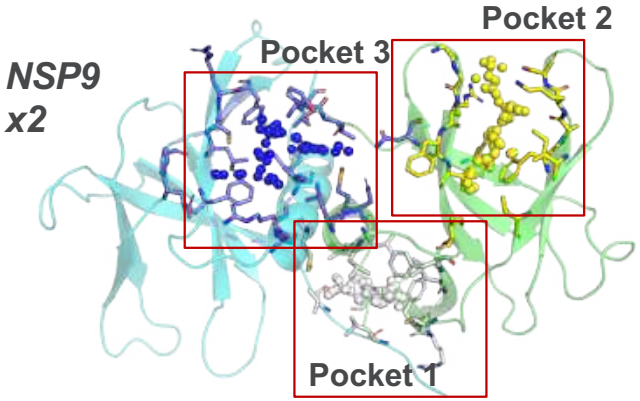
Diagram of X-ray crystallography

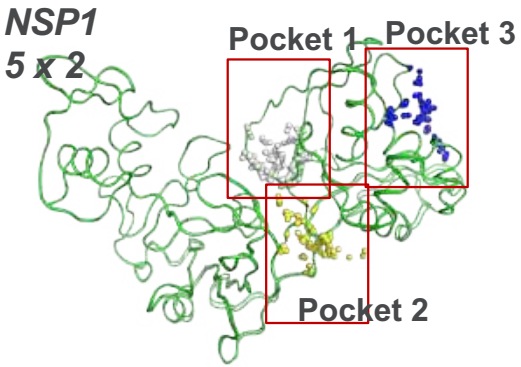
- docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex.
- It is a simple problem to understand, figure out how to fit the ligand onto the protein
- The search space in theory consists of all possible orientations and conformations of the protein paired with the ligand.

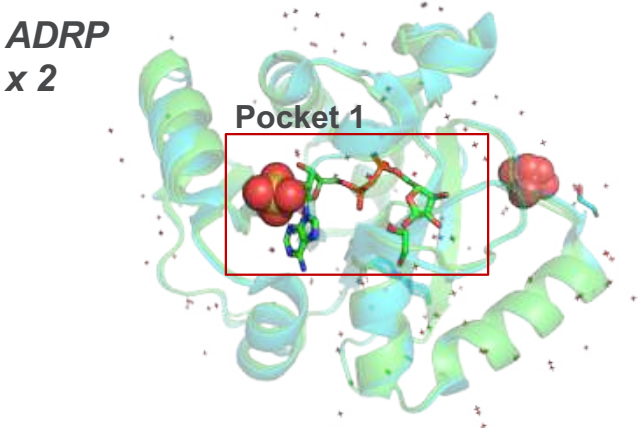
```
Inputs: molecular dataset (2D SMILES strings),
        target protein structure,
        search parameters,
        scoring function f

For ligand pose from strategy
        Result = max(Result, f(Protein pose, Ligand pose))
```

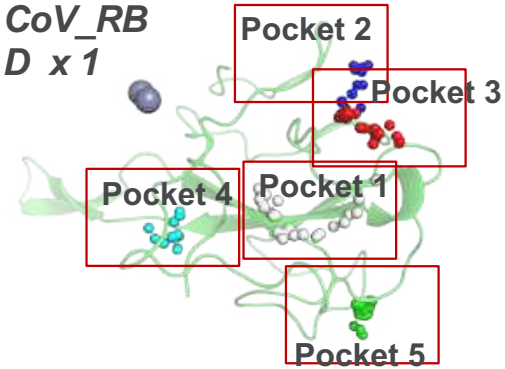


Targets and binding sites







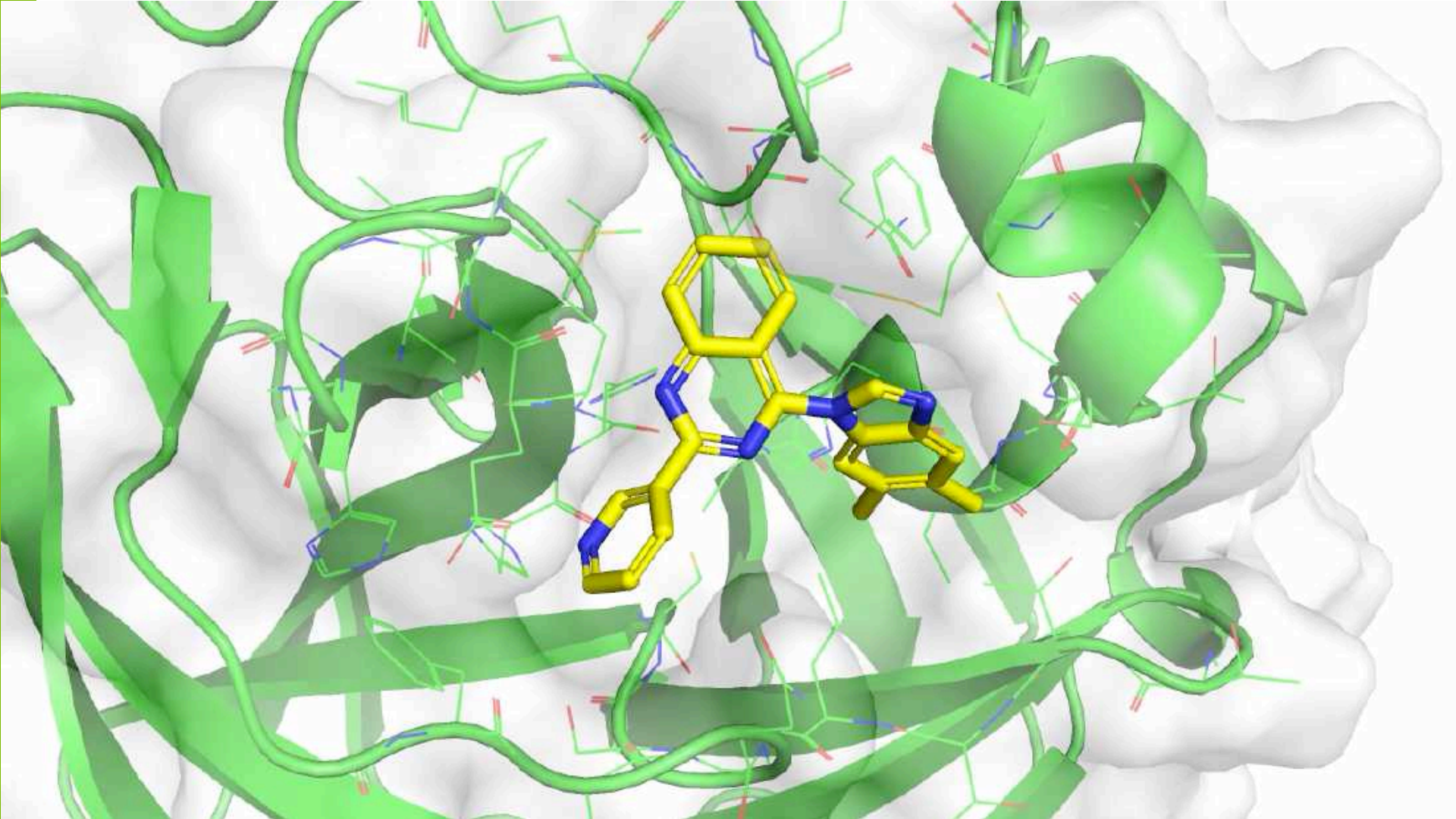


Automatic pocket detection

Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10,** 168 (2009). https://doi.org/10.1186/1471-2105-10-168

Pocket 1 :
- Score : 0.915
- Druggability Score : 0.920
- Number of Alpha Spheres : 80
- Total SASA : 16.657
- Polar SASA : 2.165
- Apolar SASA : 14.492
- Volume : 599.003
- Mean local hydrophobic density : 18.690
- Mean alpha sphere radius : 3.963
- Mean alp. sph. solvent access : 0.523
- Apolar alpha sphere proportion : 0.363
- Hydrophobicity score: 33.000
- Volume score: 3.143
- Polarity score: 4
- Charge score : 0
- Proportion of polar atoms: 39.583
- Alpha sphere density : 5.345
- Cent. of mass - Alpha Sphere max dist: 14.313
- Flexibility : 0.118

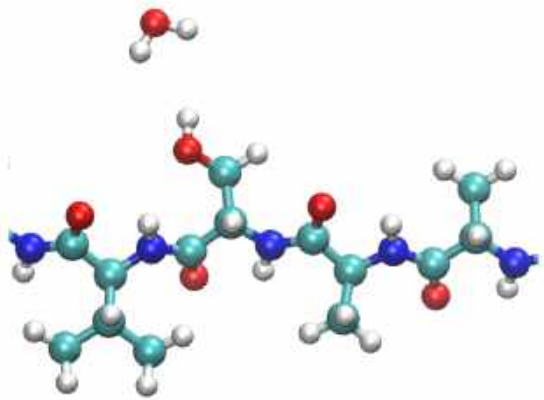
Pocket 2 :
- Score : 0.689
- Druggability Score : 0.834
- Number of Alpha Spheres : 67
- Total SASA : 8.089
- Polar SASA : 3.259
- Apolar SASA : 4.831
- Volume : 367.098
- Mean local hydrophobic density : 20.545
- Mean alpha sphere radius : 3.909
- Mean alp. sph. solvent access : 0.483
- Apolar alpha sphere proportion : 0.328
- Hydrophobicity score: 27.125
- Volume score: 2.875
- Polarity score: 3
- Charge score : 1
- Proportion of polar atoms: 40.541
- Alpha sphere density : 3.665
- Cent. of mass - Alpha Sphere max dist: 10.679
- Flexibility : 0.124



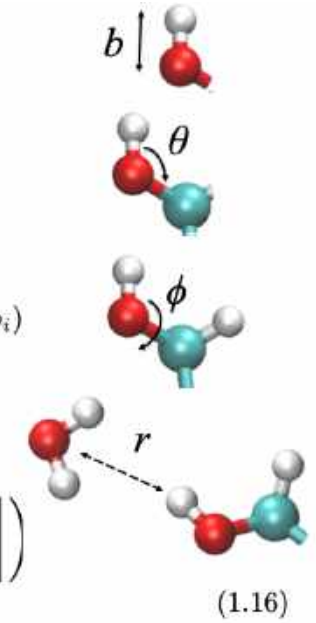
Protein-Ligand Interactions

How can we measure the attraction?

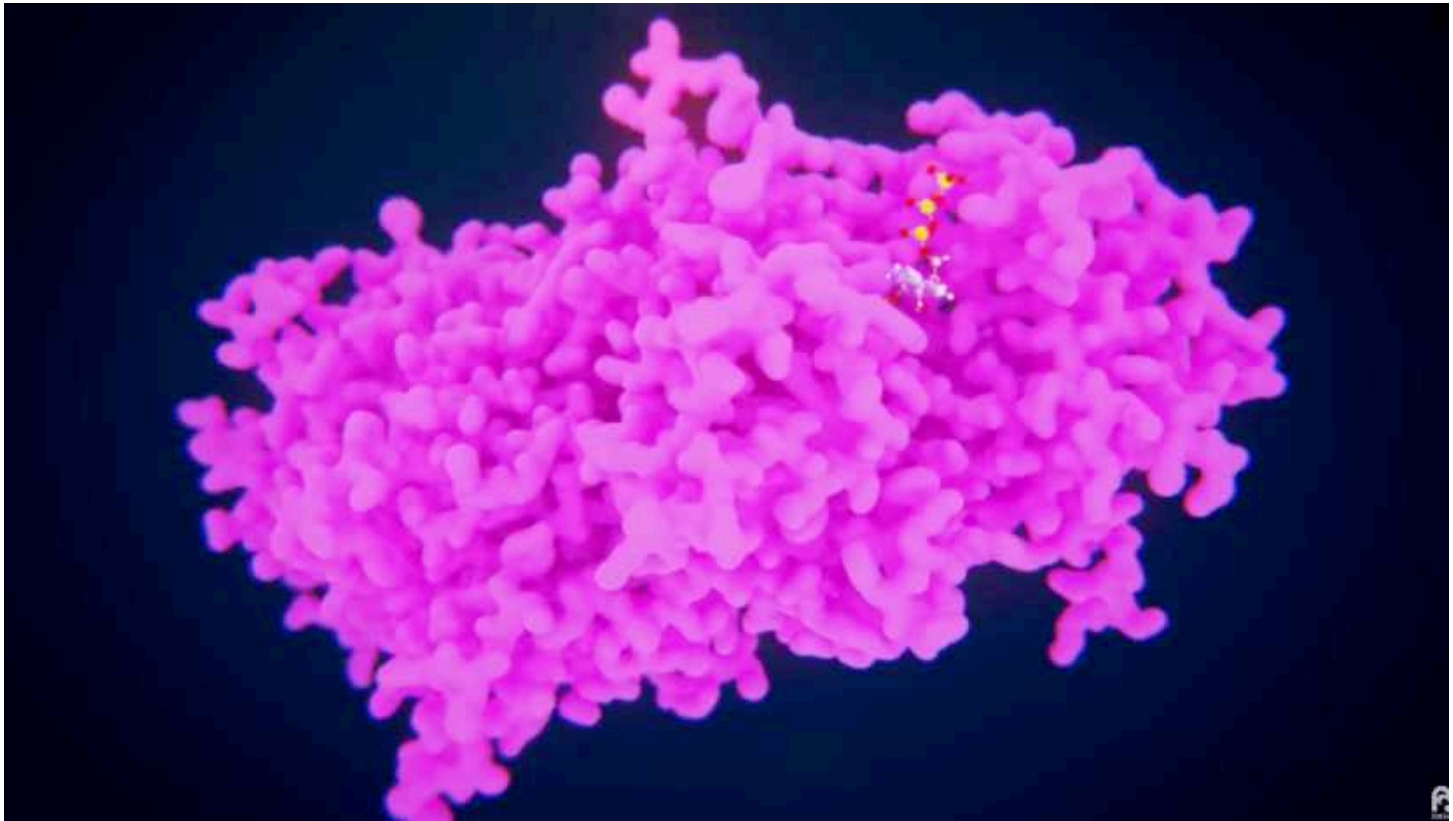
- The most realistic is quantum mechanical
- The Born-Oppenheimer (BO) approximation
- force field constructed from simple analytical and differentiable functions.



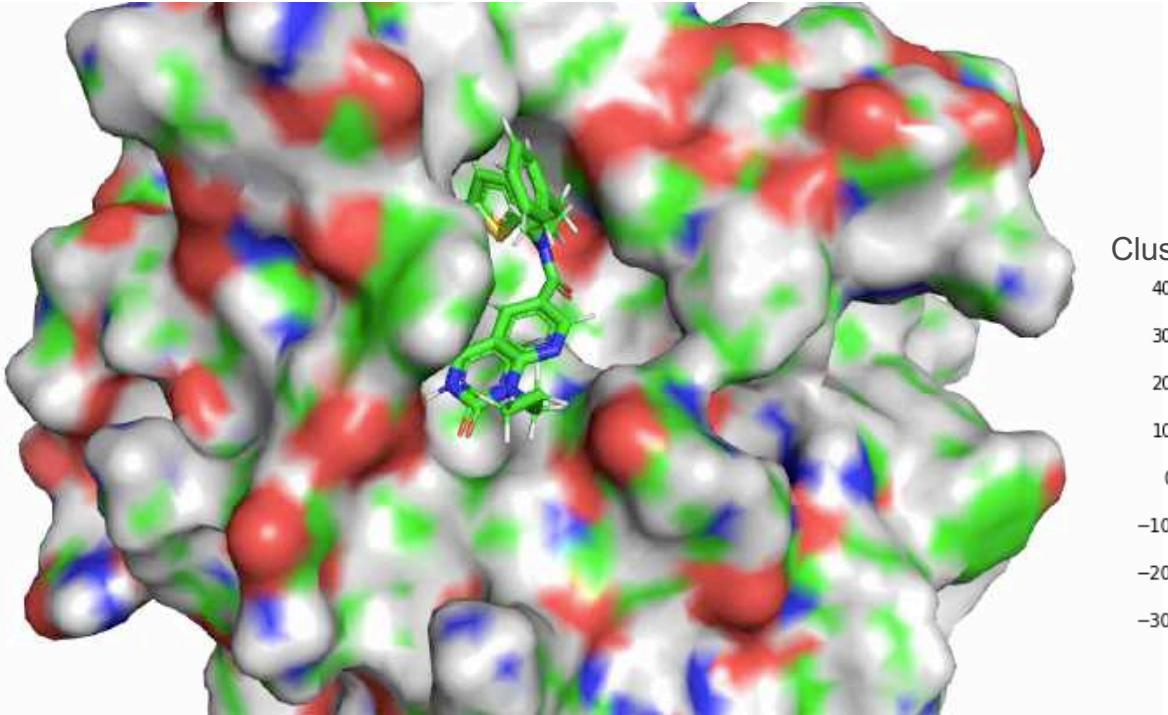
$$U(\mathbf{r}_1,\ldots,\mathbf{r}_N) = \sum_{\text{all bonds}} \frac{1}{2}K_i^{(\text{bond})}(b_i - b_{\text{eq},i})^2$$
$$+ \sum_{\text{all angles}} \frac{1}{2}K_i^{(\text{angle})}(\theta_i - \theta_{\text{eq},i})^2$$
$$+ \sum_{\text{all dihedrals}} \sum_i K_i^{(\text{dihedral})} \cos(m_i\phi_i)$$
$$+ \sum_{\text{all nonbonded pairs}} \left(\frac{q_i q_j}{\epsilon r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right) \quad (1.16)$$



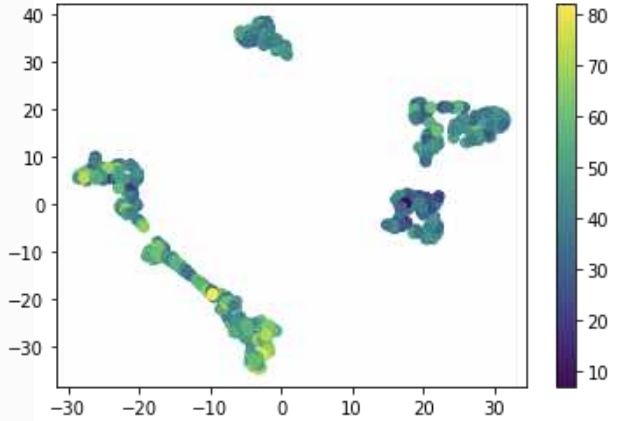
It’s easy to fall pray to Ehrlich’s idea of the magic bullet



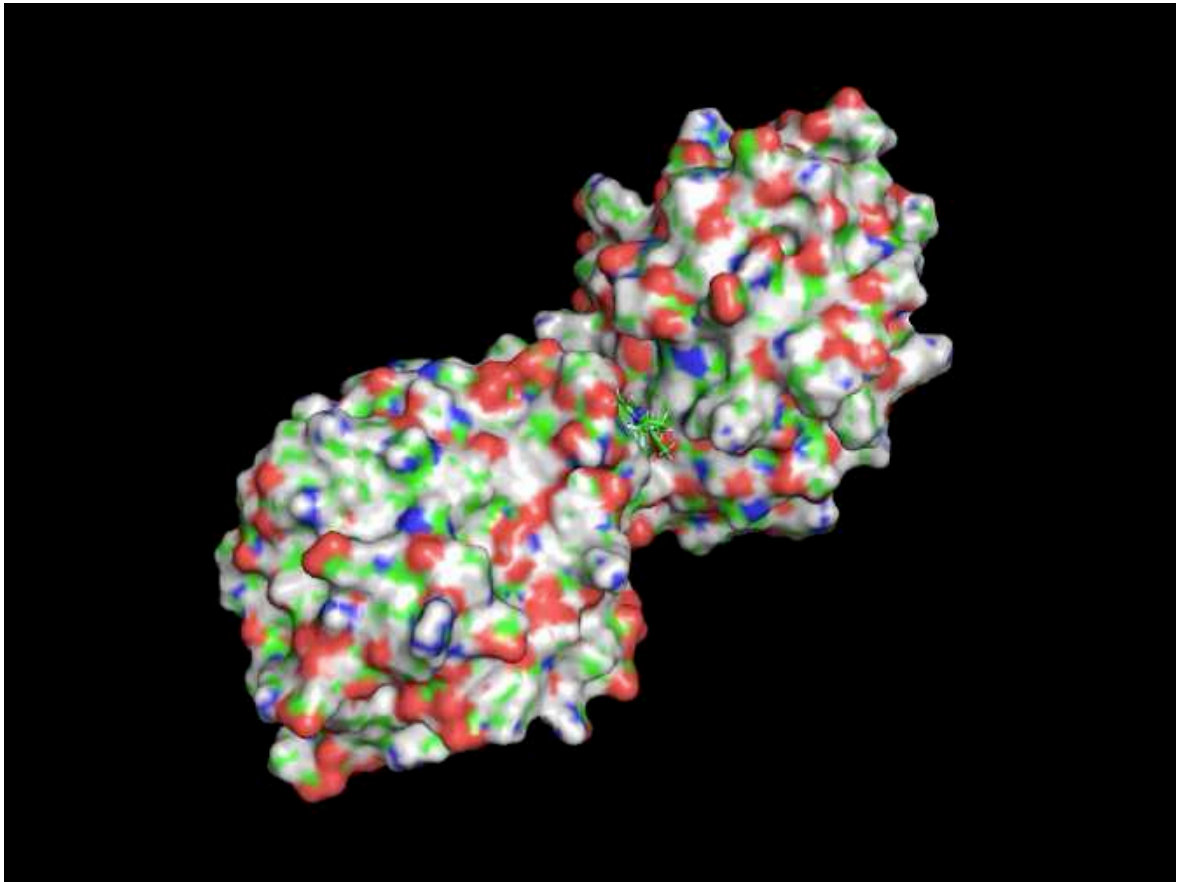
MD Simulation



Cluster states, color by protein ligand contacts

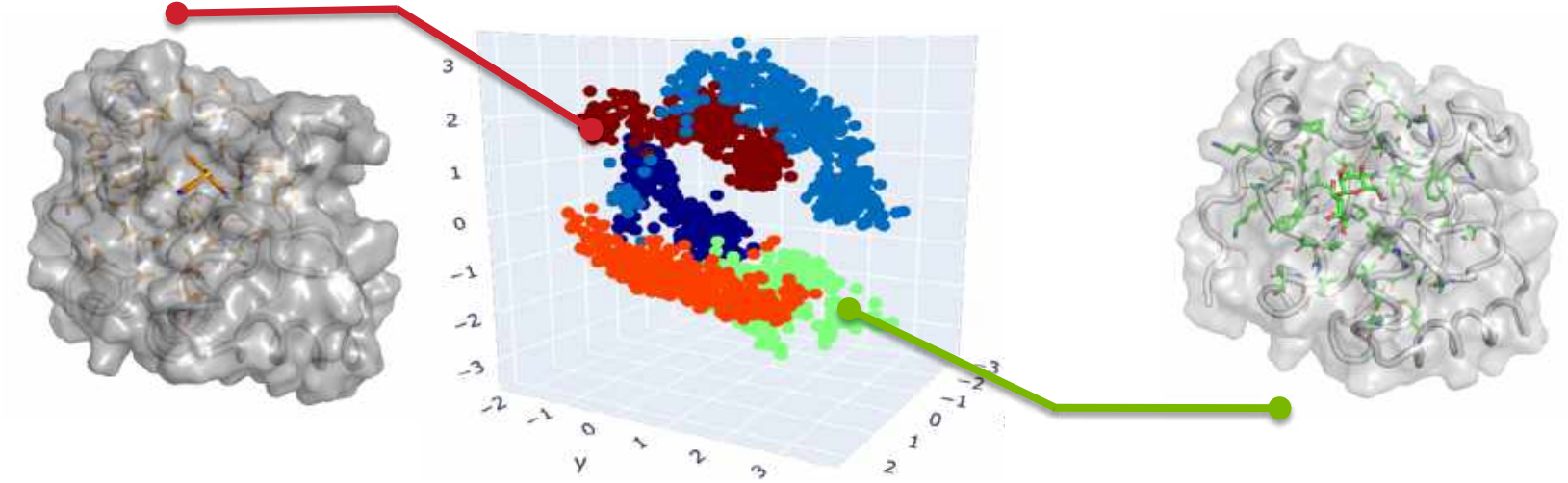


MD Simulation



H. Ma, D. Bhowmik, H. Lee, M. Trill, S. Jha, A. Ramanathan, Scalable Execution of Protein Folding with Deep Learning Driven Adaptive Molecular Simulations, ParCo 2019

Deep Drive MD



BFE workflow

How can we get simulation level accuracy of binding free energy on a dataset with over 10B ligands?

Current sample workflow

1,000,000 poses docked

Top | 2.5%

25,000 systems build and minimized

Top | 2.5%

625 systems simulated

System Components

ML Screening

~0.001 seconds per task per GPU

Structural Docking

~8 seconds per ligand per CPU core

System building and minimization

~30 seconds per ligand per GPU

MMGBSA (simulation)

~15 minutes per ligand per GPU

Goal Workflow

10,000,000,000 compounds screened with AI models

Top | 2.5%

250,000,000 poses docked

Top | 2.5%

6,250,000 systems build and minimized

Top | 2.5%

156,250 systems simulated

(that's about 12H on 1024 summit nodes)

High Throughput Docking (HT Docking)

On Demand Library Sizes

Product name	Format/Size	Descriptions	Download file
Discovery Diversity Set	10 560 compounds	High-quality diverse library of latest compounds	
Discovery Diversity Set	50 240 compounds	Top-quality diverse library of recently synthesized compounds	
Hit Locator Library	300 115 compounds	A sizable highly diverse screening set	
Phenotypic Screening Library	6 370 compounds	Special diversity set created for Phenotypic Screens	
Covalent Screening Library	32 411 compounds	Largest and most reliable source of Covalent Modifiers	

- Real Explorer* – Estimated 13B cpds
- Real – Library 1.2B cpds
- Diverse *REAL* drug-like, 15M cpds
- *REAL* lead-like, 868M cpds
- *REAL* 350/3 lead-like, 378M cpds

LETTER

https://doi.org/10.1038/s41586-019-1540-5

Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia[1], Allyson Lynch[1], Yuheng Huang[1], Matthew Danielson[1], Immaculate Lang'at[1], Alexander Milder[1], Aaron E. Ruby[1], Hao Wang[1], Sorelle A. Friedler[2]*, Alexander J. Norquist[1]* & Joshua Schrier[1,2]*

- Use about 115k building blocks

"Machine-learning models that we train on a smaller randomized reaction dataset outperform models trained on larger human-selected reaction datasets, demonstrating the importance of identifying and addressing anthropogenic biases in scientific data."

*synthesis time is 3-4 weeks with an average success rate of over 80%

HT Docking

Article | Published: 06 February 2019

Ultra-large library docking for discovering new chemotypes

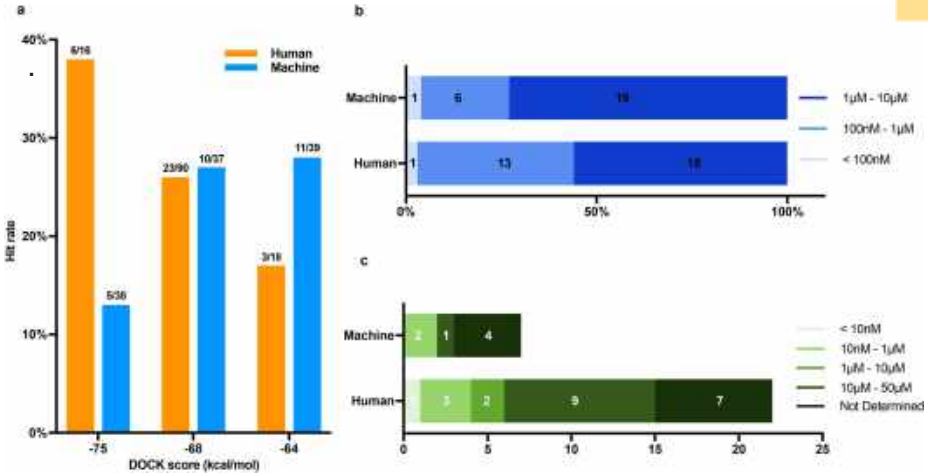
Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet ✉, Bryan L. Roth ✉ & John J. Irwin ✉

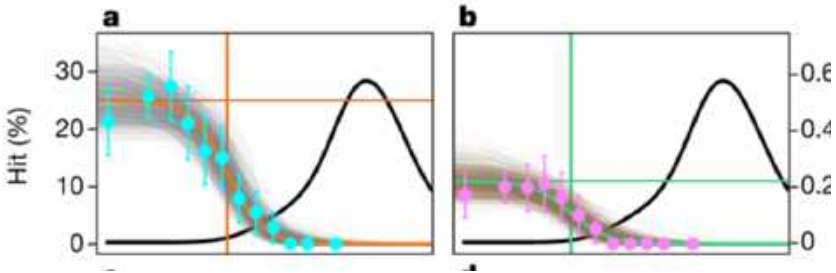
Nature 566, 224-229(2019) | Cite this article

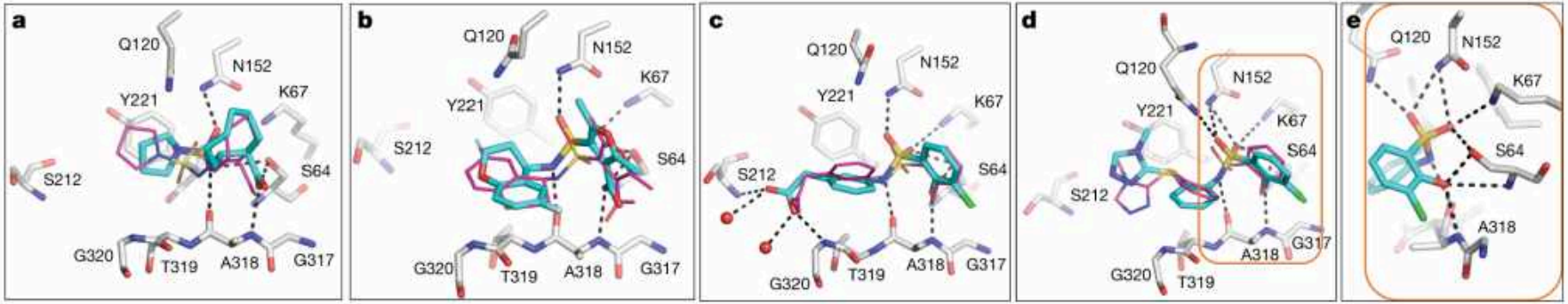
37k Accesses | 60 Citations | 273 Altmetric | Metrics

- Lyu *et al.* found that it was essential to screen the full library to discover the most biologically active compounds
- Human inspection was somewhat similar
- the molecules prioritized by human inspection typically had better affinities: 44% of these were submicromolar, which was true of only 27% of those prioritized by docking score alone.

1. Dock 99 million molecules
   i. Average of 4,054 orientations
   ii. Average of 280 conformers
2. Cluster top-ranked 1 million
   i. Remove similarity compounds to known inhibitors
   ii. Reducing redundancy
3. Fifty-one top-ranking molecules
   i. 44 (86%) were successfully synthesized
   ii. 5 compounds measurably inhibited, 11% hit rate
4. Lead Optimization
   i. To optimize the five initial hits, we chose 90 well-scoring analogues from within the make-on-demand library
   ii. Over half were active on testing, improving the affinity of each of the 5 hits by 3- to 29-fold







Crystal structures of the inhibitors (carbons in cyan) overlaid with their docking predictions (magenta). AmpC carbon atoms are shown in grey, oxygens in red, nitrogens in blue, sulfurs in yellow, chlorides in green and fluorides in light blue. Hydrogen bonds are shown as black dashed lines. **a**–**d**, AmpC in complex with ZINC547933290 (**a**; Protein Data Bank (PDB) 6DPZ, r.m.s.d. = 1.3 Å) and 275579920 (**b**; PDB 6DPY, r.m.s.d. = 1.2 Å for the warhead), the 1.3 μM inhibitor 339204163 (PDB 6DPX; r.m.s.d. = 0.98 Å) and its 77 nM analogue 549719643 (**d**; PDB 6DPT, r.m.s.d. = 1.52 Å). **e**, Close-up of the 549719643 phenolate in the oxyanion hole. Extended Data Fig. 4 shows the electron densities.

RADICAL-Cybertools & COVID-19

http://radical.rutgers.edu/

- "The throughput averaged 1s per library compound" – Lyu paper
- "1,513,728,000" CPU seconds to compute on enamine real database (1.4B)
- We're running around 0.5s per library compound, working on bringing that up

	Scientific Performance	Computational Performance	Platforms Used
Workflow-0 HT Docking	• 300K/h ligands on Frontera (TACC) on 4 nodes • 50K/h on Theta (ANL) on 128 nodes	Linear weak scaling up to 512 nodes (TBC)	• Frontera (TACC) • Comet (SDSC) • Theta (ANL) • Stampede2 (TACC)
Workflow-1 Automated System building and MMGBSA	• Andre to provide	• Typical at 128 GPUs • Planned at 1024 GPUs	• Summit (OLCF) • Longhorn (TACC) TBC
Workflow-2 (DeepDrive MD)	• O(10)-O(100) faster sampling compared to Anton trajectory for BBA, VHP proteins • For COVID19 assembly proteins TBC	• Typical 144 GPUs (128 MD + 16 CVAE) • Up to 6 concurrent workflows	• Summit (OLCF) • Longhorn (TACC) TBC

Aurora: HPC and AI

Argonne NATIONAL LABORATORY

> ExaFlops/s for HPC
>> Exaops/s for AI

intel

CRAY

Architecture supports three types of computing

- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

Accelerating with ML

Quick bit on drug discovery philosophy

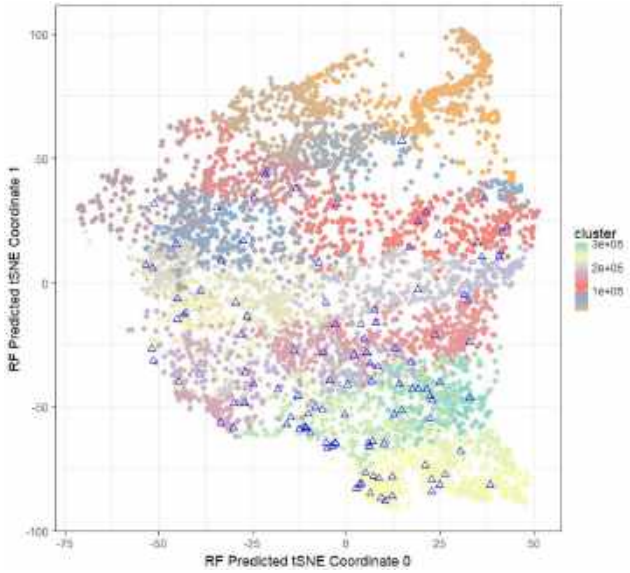
I see two camps out here

Brute-force

- Need some sort of oracle to filter space

Scaffold hoping

- Once a good hit is found, sample densely around it

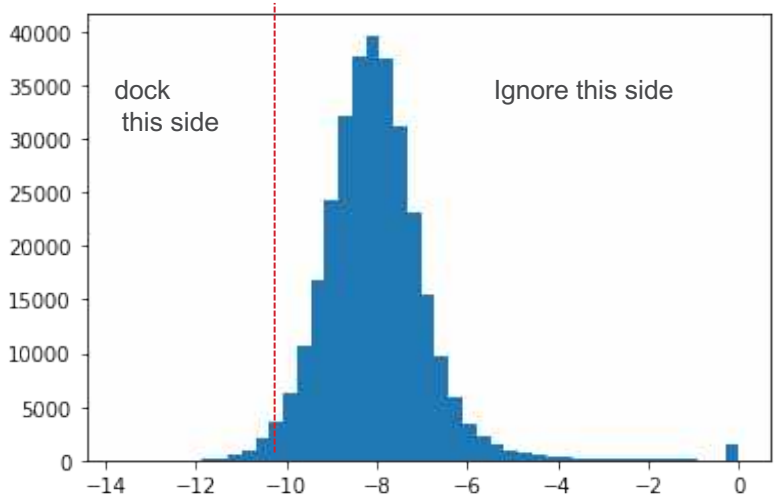


Accelerating with ML

Recall, when you're docking you are exhaustively finding a pose, wasting CPU cycles and IO

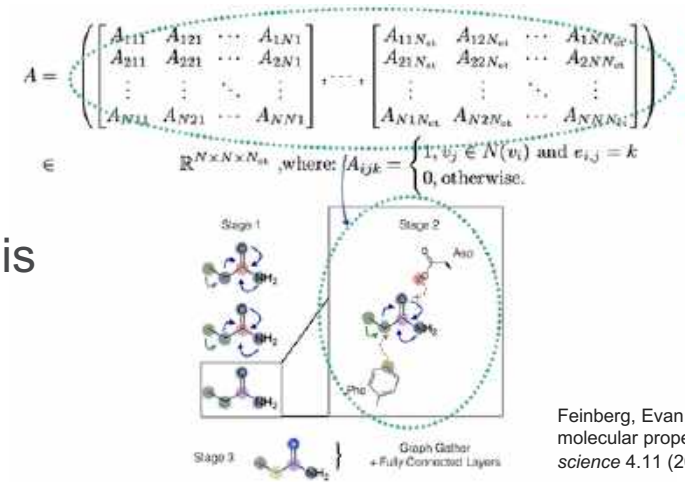
	smiles	ADRP-ADPR_pocket1_dock
2531	Brc1ccccc1c1nnc(o1)Cn1nc(c(c1C)[N+](=O)[O-])C	-3.503724
196791	N#Cc1ccccc1OCc1csc(n1)c1ncccn1	-3.776353
143928	Cc1cc(C(=O)C)c(c(c1)C(=O)C)OCc1ccc(cc1)Cl	-3.148670
9558	CC(=O)Nc1ccc(cc1)SCc1[nH]c(=O)c2c(n1)c1ccccc1o2	-3.383467
290589	O=c1[nH]c(=O)c2c(n1)n(C[C@@H]([C@@H]([C@@H](CO...	0.000000
...	...	...
194200	N#Cc1ccc(cc1)OCc1noc(n1)C	-5.183101
101391	COCCNc1nnc(s1)SCC(=O)c1cc(n(c1C)C1CC1)C	-2.622271
104532	CO[C@@H]1CN(C[C@H]1c1nnn(c1)C)C(=O)c1cc(O)nc(c...	-2.766863
78848	CN(CC(=O)Nc1cc(F)cc(c1)F)CCC(=O)O.Cl	-3.856561
262213	O=C(Nc1ccnn1Cc1ccc(cc1)Cl)CCc1c(C)noc1C	-2.428766

310404 rows x 2 columns

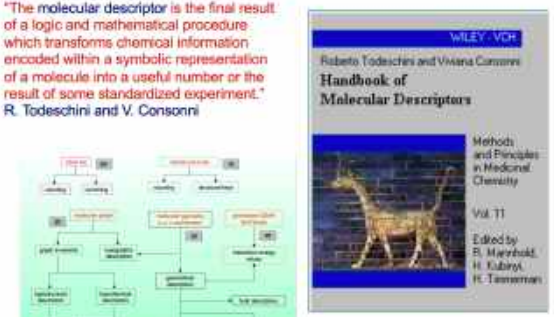


Build an ML model that:

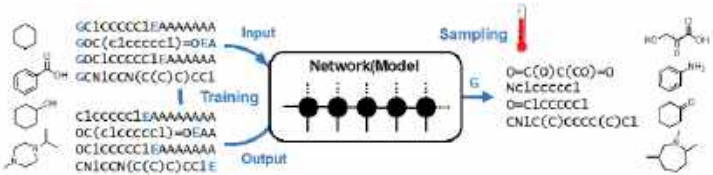
- Takes a ligand in and featurize it
- Outputs some number
- Take that number as this is a good compound go ahead and dock or bad compounds don't waste your time



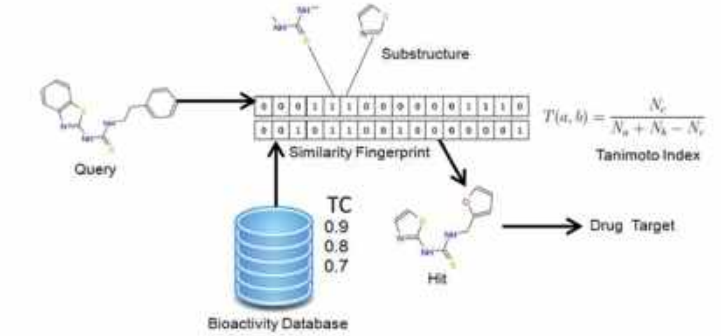
"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."
R. Todeschini and V. Consonni



Feinberg, Evan N., et al. "Potentialnet for molecular property prediction." *ACS central science* 4.11 (2018): 1520-1530.



Gupta, Anvita, et al. "Generative recurrent networks for de novo drug design." Molecular informatics 37.1-2 (2018): 1700111.



Example simple model

Use a random forest model on cheminformatic descriptors

```
In [279]: df = df[['smiles','ADRP-ADPR_pocket1_dock']].sample(frac=0.01)
```

```
In [*]: from mordred import descriptors, Calculator
        from rdkit import Chem

        mols  = [Chem.MolFromSmiles(df.iloc[s,0]) for s in range(df.shape[0])]
        calc = Calculator(descriptors, ignore_3D=True)
        df_des = calc.pandas(mols)
```

```
 63%|██████    | 1945/3104 [01:26<03:15,  5.93it/s]
```

	ABC	ABCGG	nAcid	nBase	nAromAtom	nAromBond	nAtom	nHeavyAtom	nSpiro	nBridgehead	...	SRW09	SRW10	TSRW10
0	17.499447	14.658900	0.0	0.0	11.0	11.0	32.0	22.0	0.0	0.0	...	6.861711	10.242065	70.806108
1	18.522676	14.458540	0.0	0.0	11.0	11.0	47.0	24.0	0.0	0.0	...	6.580639	9.804385	71.515122
2	25.650763	18.828555	0.0	0.0	24.0	26.0	52.0	32.0	0.0	0.0	...	7.423568	10.447932	84.148513
3	19.226790	15.646696	0.0	0.0	10.0	11.0	50.0	25.0	0.0	0.0	...	6.259581	10.115570	72.797457
4	17.331590	14.703517	0.0	0.0	11.0	11.0	44.0	22.0	0.0	0.0	...	6.605298	10.209832	70.298555

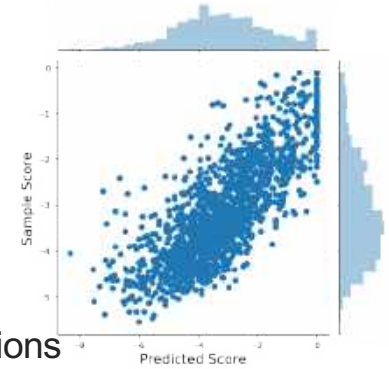
Generate Descriptors

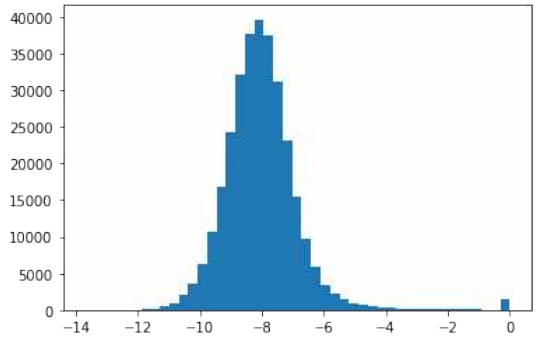
```
In [308]: from sklearn.ensemble import RandomForestRegressor
          from sklearn.model_selection import cross_val_score

          rf = RandomForestRegressor()
          scores = cross_val_score(rf,X,y,cv=5)
          print('r2_score: ' + str(np.mean(scores)) + " +/ " + str(np.std(scores)))
```

```
r2_score: 0.6246748341262527 +/ 0.01401545464858188
```

Build a Model

Analyzing the models

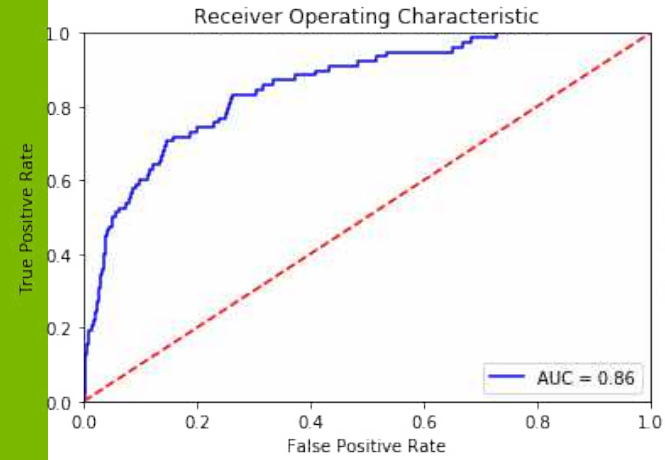


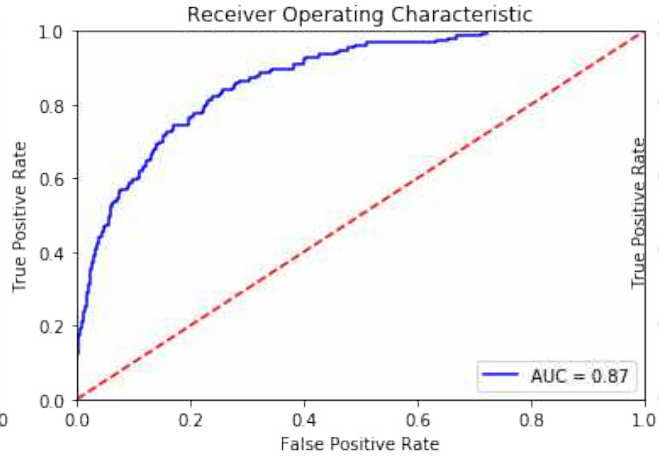


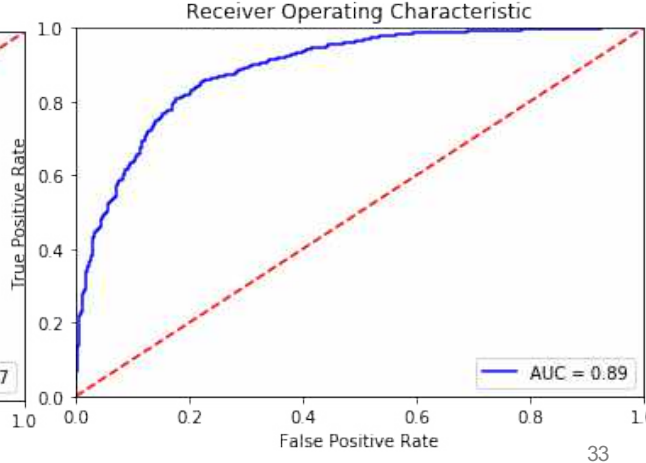
- Metrics become a difficult task.
- Suppose just want the top 5%, we can bin the predictions that come out of the model

[[1424 50] [50 28]]

1424	50
50	28





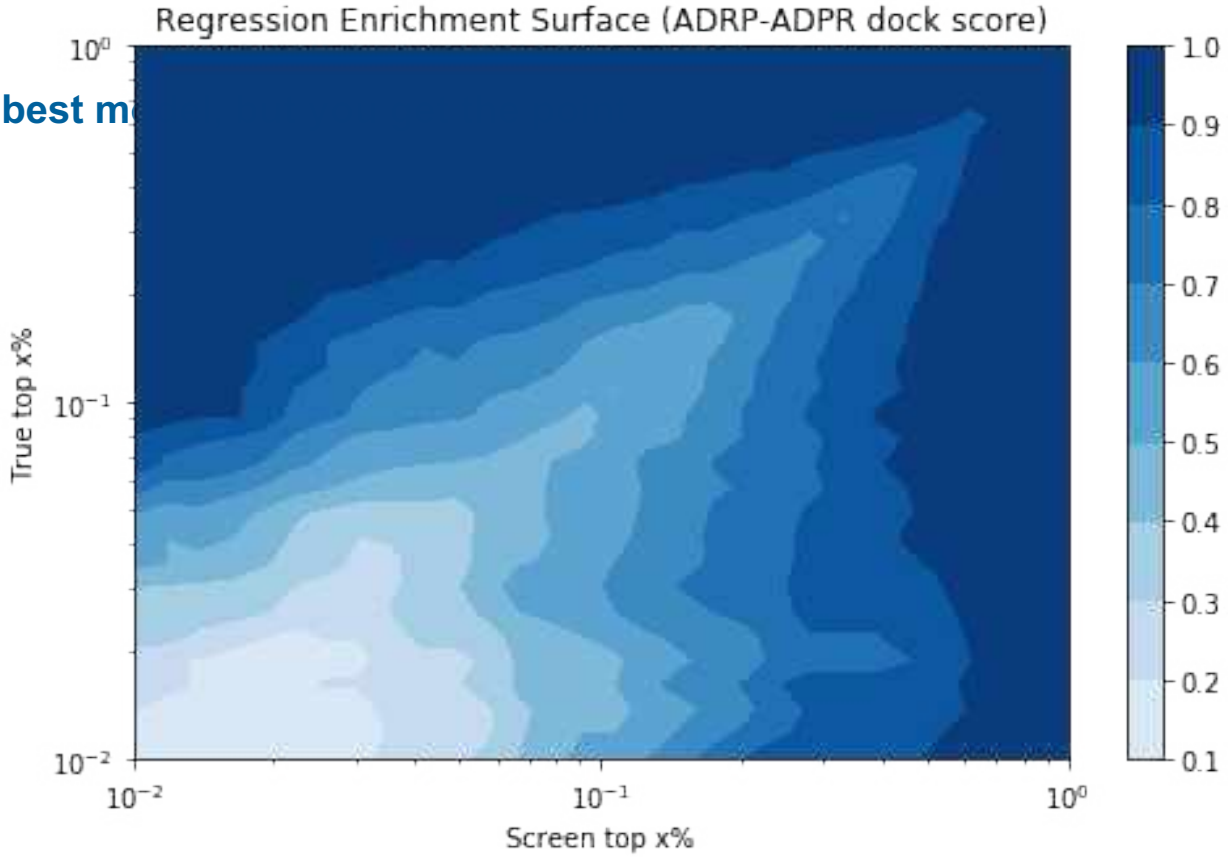


- We need two things to decide on a cut-off. We need to know how many downstream tasks we can afford to dock, how many of the top leads we want to find, and what our tolerance is for missing some.

- Let's compute it:
- For leads_desired in top [1%, 10%, 50%]
  For number_able_to_dock in predicted top [1%, 10%, 50%]
    score[leads_desired][number_able_to_dock] =
        how many of my top predictions I was able to dock found the desired leads?

  Now let's just look and see where in that grid we were able to capture the most part of the leads

Clearly not the best m



REAL database, 1.2Bn cpds

I estimate only 1M of these will be a solid interesting poses

Ingest 2048 compounds per second per GPU

2 GPUs

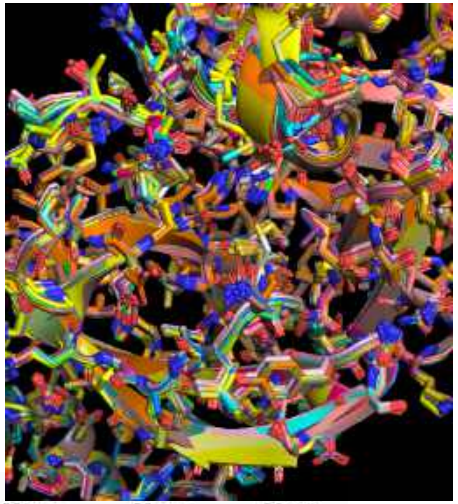
Models need to perfectly capture the top 1.36% scoring compounds

Pass through 1 ligands per second per core

Dock software

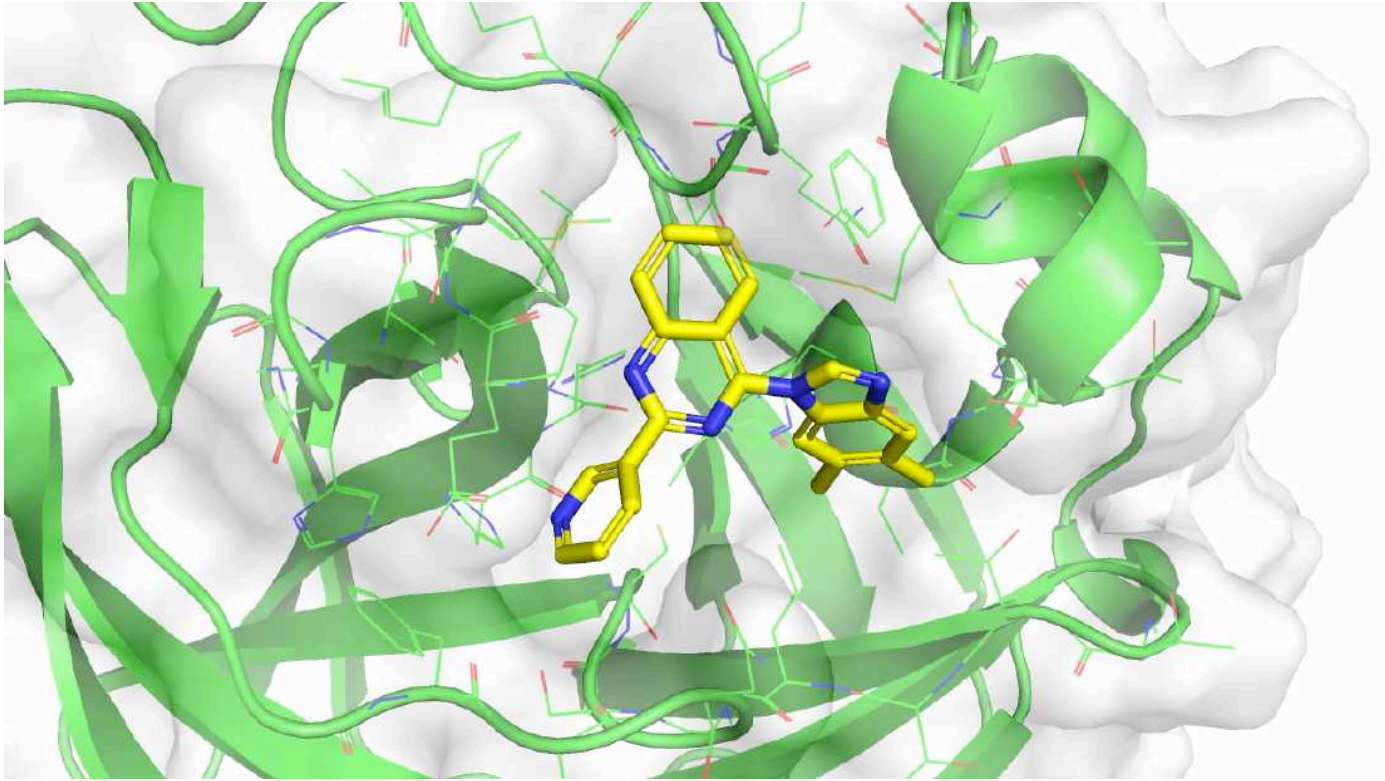
56 Cores

Ensemble Docking Strategy



	smiles	Mpro-x0072_dock	Mpro-x0104_dock	Mpro-x0107_dock	Mpro-x0161_dock	Mpro-x0195_dock	Mpro-x0305_dock	Mpro-x0354_dock
190206	N#Cc1[nH]cc(c1)C(=O)NC1CCCN(C1=O)Cc1ccc(cc1F)F	-7.610964	-9.077101	-6.082466	-7.187638	-3.866065	-8.221134	-8.407152
183577	Fc1ccc(cc1)n1nc(c2c1CCCC2)C(=O)Nc1cccc2c1nccc2	-7.651024	-8.650168	-6.455404	-9.196528	-2.535620	-7.974315	-9.516221
20569	CC(OCC1OCCN(C1)C(=O)Nc1cc(C)nnc1N(C)C)C	-6.672259	-6.808953	-4.886088	-7.568527	-8.250710	-7.454492	-7.919614
51863	CCN1CCN(CC1c1nccn1C)C(=O)c1ccoc1C	-7.782367	-7.818557	-7.168431	-8.478214	-6.264867	-8.653809	-7.940884
61669	CCOc1ccc(cc1)S(=O)(=O)Nc1ccccc1C(=O)NCc1ccccc1	-6.892548	-7.140425	-7.680151	-8.484144	-5.973836	-8.021129	-6.789618
...	...	...	...	...	...	...	...	...
47393	CCN(C1CCS(=O)(=O)C1)Cc1ccc(cc1)Cl	-7.675440	-8.849446	-7.132509	-8.277190	-7.847459	-8.122643	-7.028652
3998	C=CCN(C(=O)Cc1csc(n1)c1cccnc1)Cc1ccc(cc1)OC	-7.226417	-8.490604	-5.619135	-7.982813	-6.734933	-7.797452	-7.357950
307188	Oc1ccc(nn1)C1CCCN1Cc1nnc(n1C1CC1)C	-8.533868	-8.651809	-7.508603	-8.130071	-8.237130	-9.066695	-8.418077
107597	COc1cc(CCC(=O)Nc2nnc(s2)C(C)C)ccc1OC	-6.423342	-7.337890	-5.327817	-7.572044	-7.347135	-7.294445	-6.882806
258487	O=C(Nc1cccc(c1)c1ccc(=O)[nH]n1)NCCc1ccc(cc1)C	-6.501783	-8.285757	-5.552805	-7.929476	-5.661306	-8.398631	-6.669324

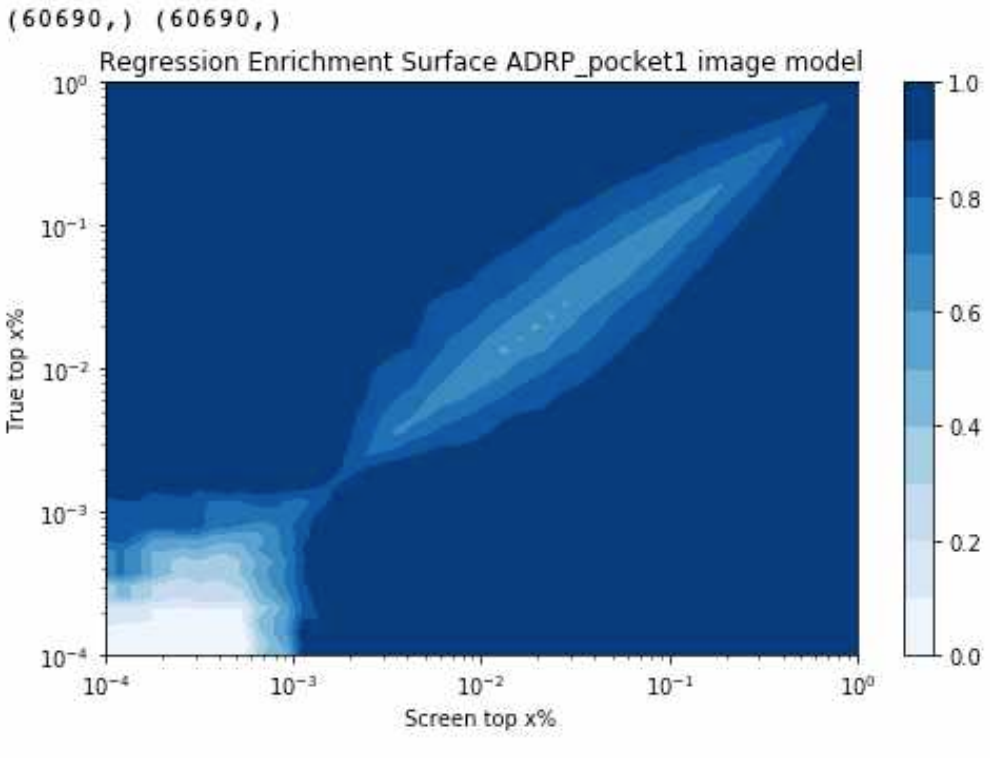
Example results from an ensemble docking run on Mpro from COVID-19



Filtering space with ML Models

130X speed up over non-ML approach

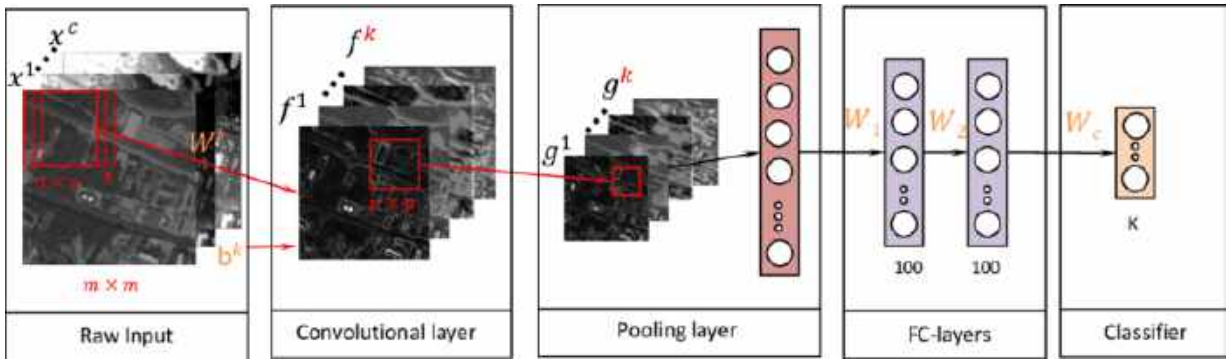
- ResNet-like model trained using 2D image depictions of molecules.
- Model shows we can retrieve the top 0.1% of dock scorers while only screening the top 0.5% of the database.
- Rather than docking 1 billion compounds to get the answer, we only need to dock 5M compounds.
- At the end, we have the same computed structures as the non-ML group, we just used 200x less CPU compute.
- 8,192 ligands per node/s can be inferenced on a GPU while only 56 ligands can be docked on the corresponding CPU node/s



Images as model features

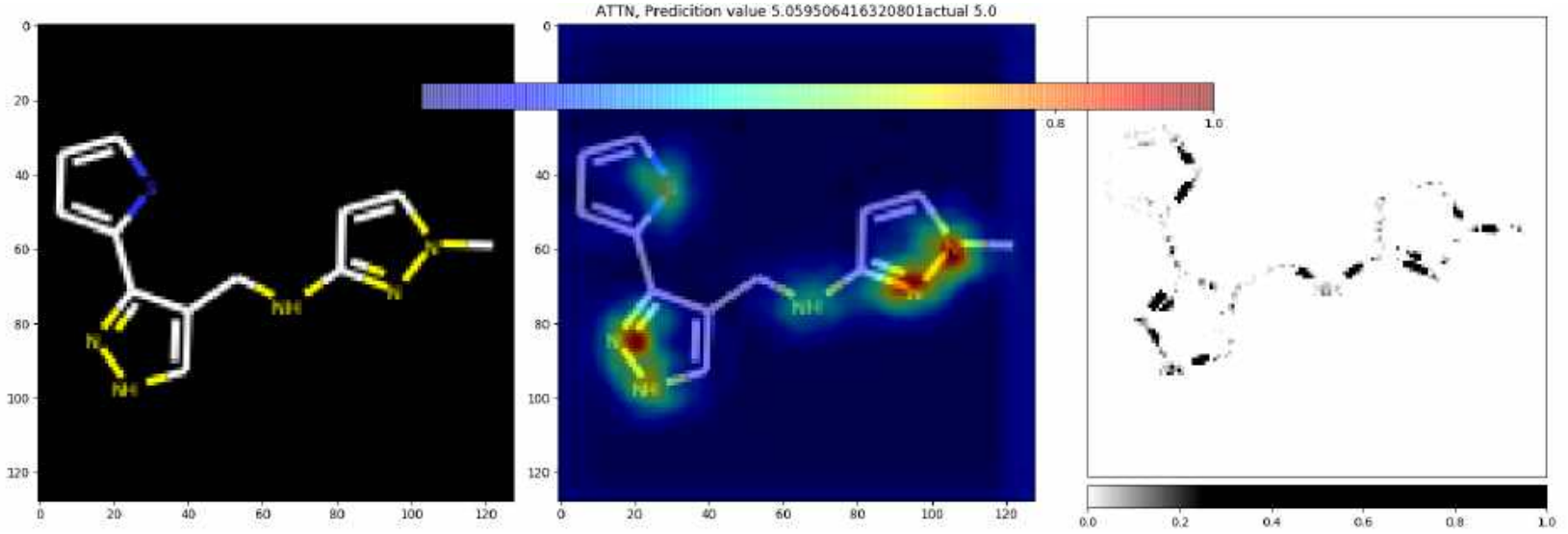
- Taking from the success of image convolutional models from computer vision tasks:
  - Can we use the image directly instead of computing descriptors or fingerprints to input the drug into our model?

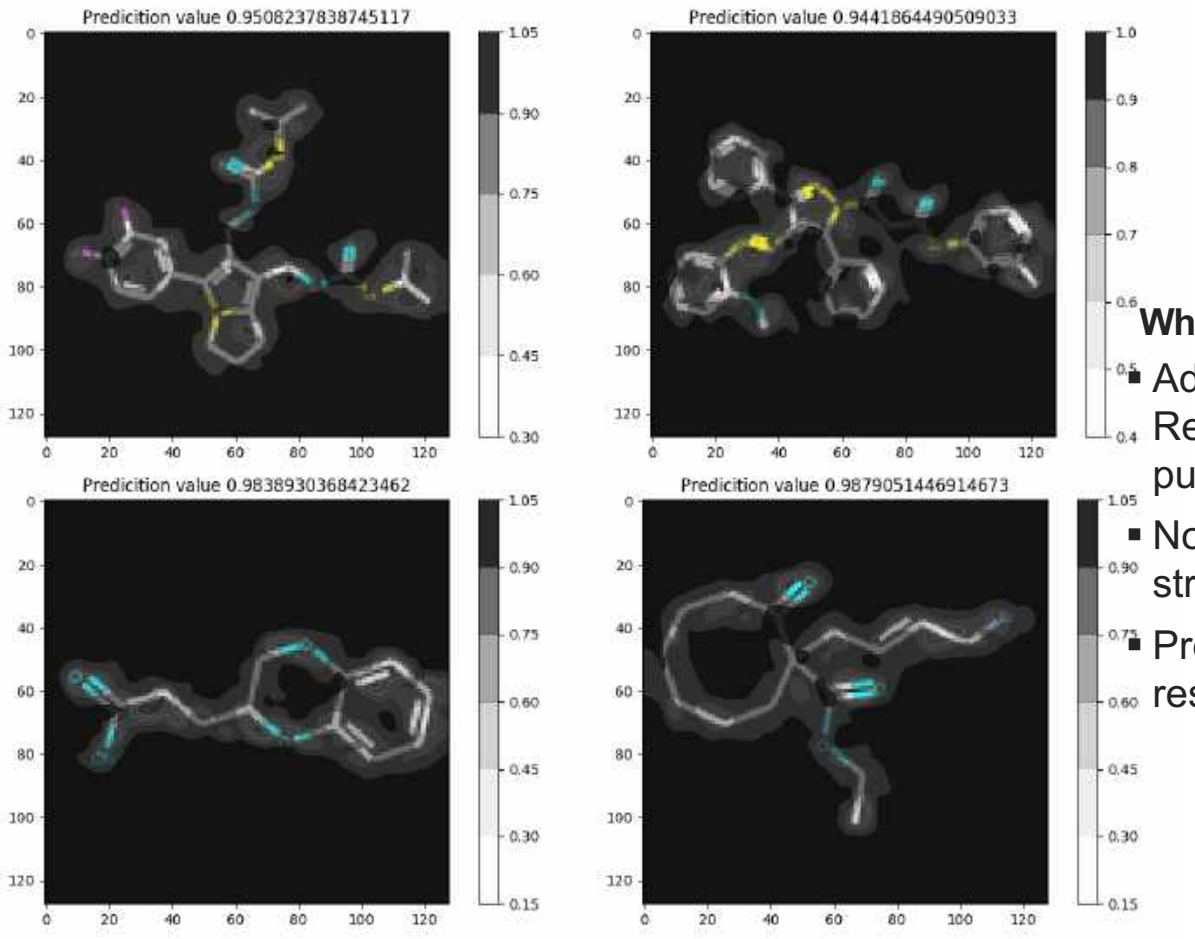
- In practice, imagine taking every smile string and generating the 2D depiction. We use that instead of the string, or the descriptors.



Example

- Suppose we wanted a model to count hydrogen bond acceptors. We would use this image (left) as input to our model. The model recognizes and counts the h-acceptors.

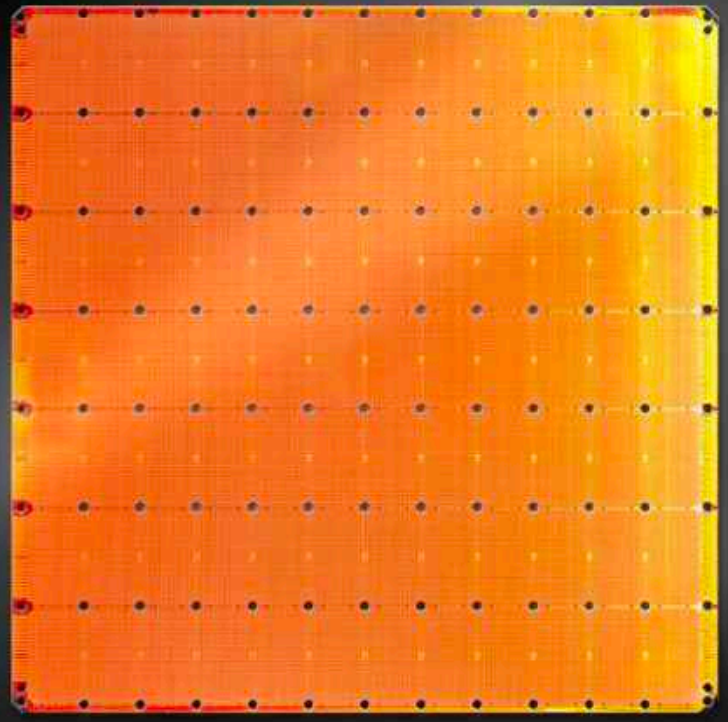




What's going on with images?

- Added simple attention in a middle ResNet layer through a large conv filter put through softmax (globally)
- Not based on cells, due to tower structure
- Prediction value is area under the dose response curve

Cerebras Wafer Scale Engine



Cerebras WSE
1.2 Trillion Transistors
46,225 mm² Silicon

Largest Chip Ever Built

- 46,225 mm^2 silicon
- 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 Gigabytes of On-chip Memory
- 9 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth
- TSMC 16nm process

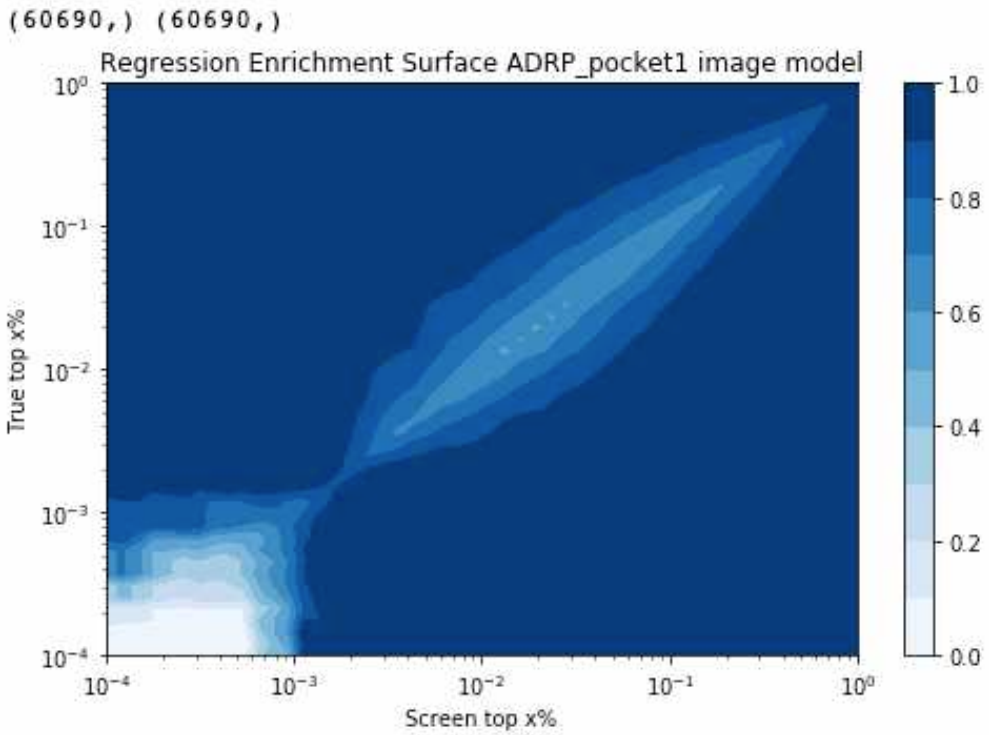
Largest GPU
21.1 Billion Transistors
815 mm² Silicon

~~130X speed up over non-ML approach~~

990X speed up with successful hardware accelerators

Filtering space with ML Models

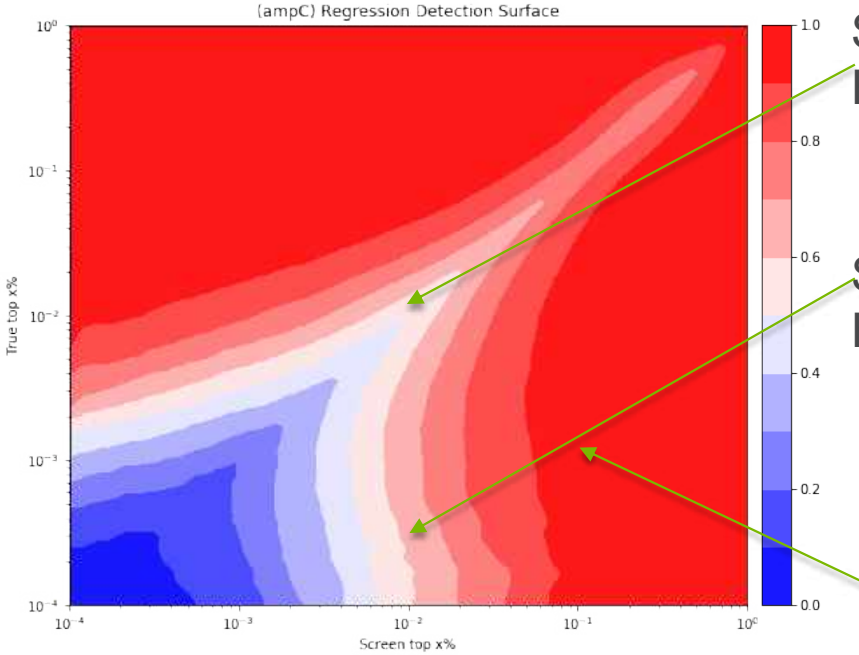
- ResNet-like model trained using 2D imag depictions of molecules.
- Model shows we can retrieve the top 0.1 of dock scorers while only screening the 0.5% of the database.
- Rather than docking 1 billion compounds get the answer, we only need to dock 5M compounds.
- At the end, we have the same computed structures as the non-ML group, we just used 200x less CPU compute.
- 102,400 ligands per node/s can be inferenced on a CS-1 while only 56 ligan can be docked on the corresponding CP node/s



What if we replace the need to simulate every molecule?

Replicating Lyu et al. Giga-Docking with 200x less CPU compute

Trained message-passing network with 500K ampC



Screen 1% of molecules, you'll Have 50% of the true top 1%

Screen 1% of molecules, you'll Have 70% of the true top 0.05%

Screen 10% of molecules, get all of the top 0.1%

*preliminary work, first approximation of a good model

Acceleration for filtering models

- Given a function you want to filter F, how can we determine what kind of acceleration you can get?
- Two components
  - Model enrichment: how many samples can you save running on F?
  - Model speed: how long does the model take to run compared to F?

$$- \frac{N * FSpeed}{N * ModelSpeed + EF * N * FSpeed}$$

Different approaches to ML and Docking

Look at ML through some lenses

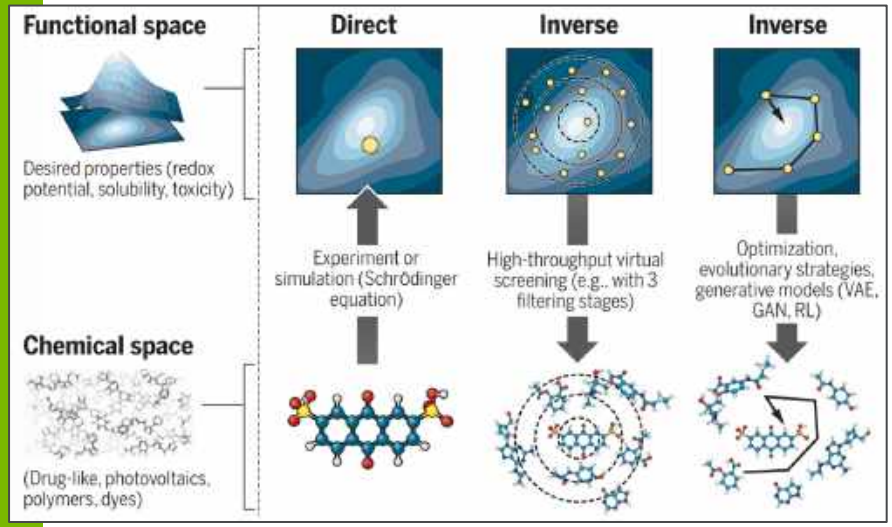
Surrogate Modeling → How to utilize AI for replacing our applications for speed?

Library generation → How can AI expand the space of possibilities?

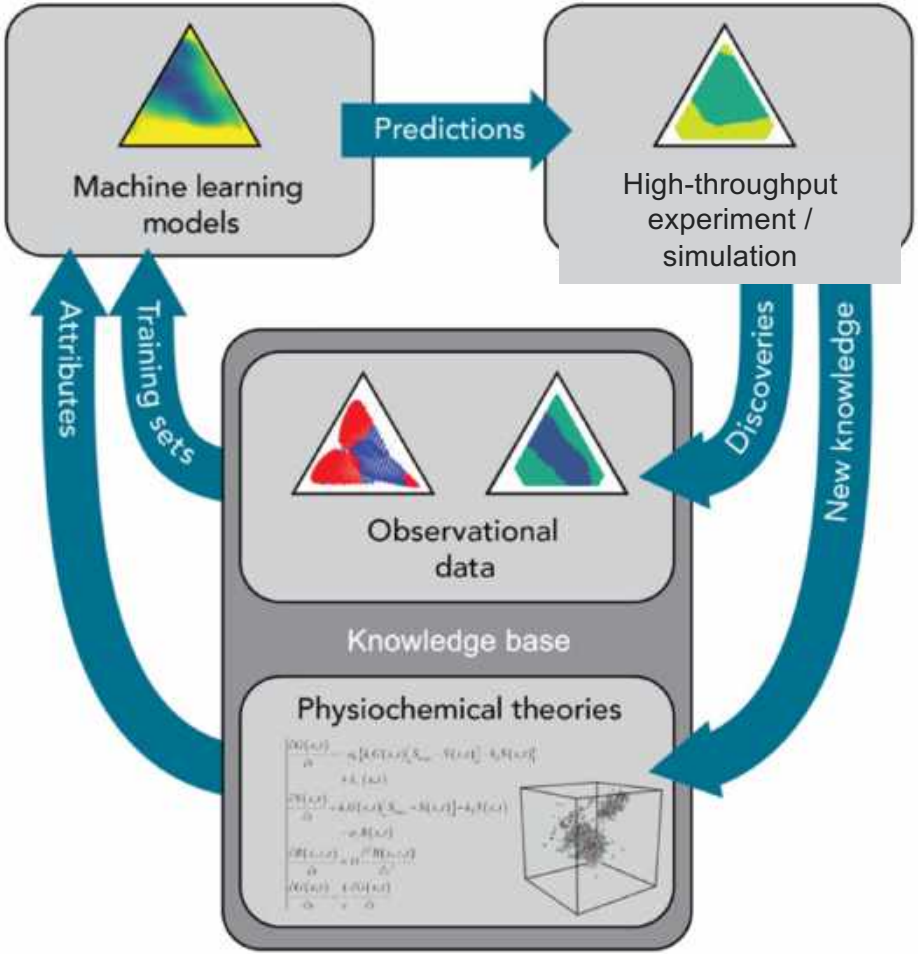
ML-driven optimization → How can AI do these tasks simultaneously?

AI can enable a new era of discovery based on automated search

- Expand knowledge: e.g., machine reading
- Autoencoders to navigate chemical and functional spaces
- Active learning to choose next experiment or simulation
- Reinforcement learning to guide experiments and/or simulations

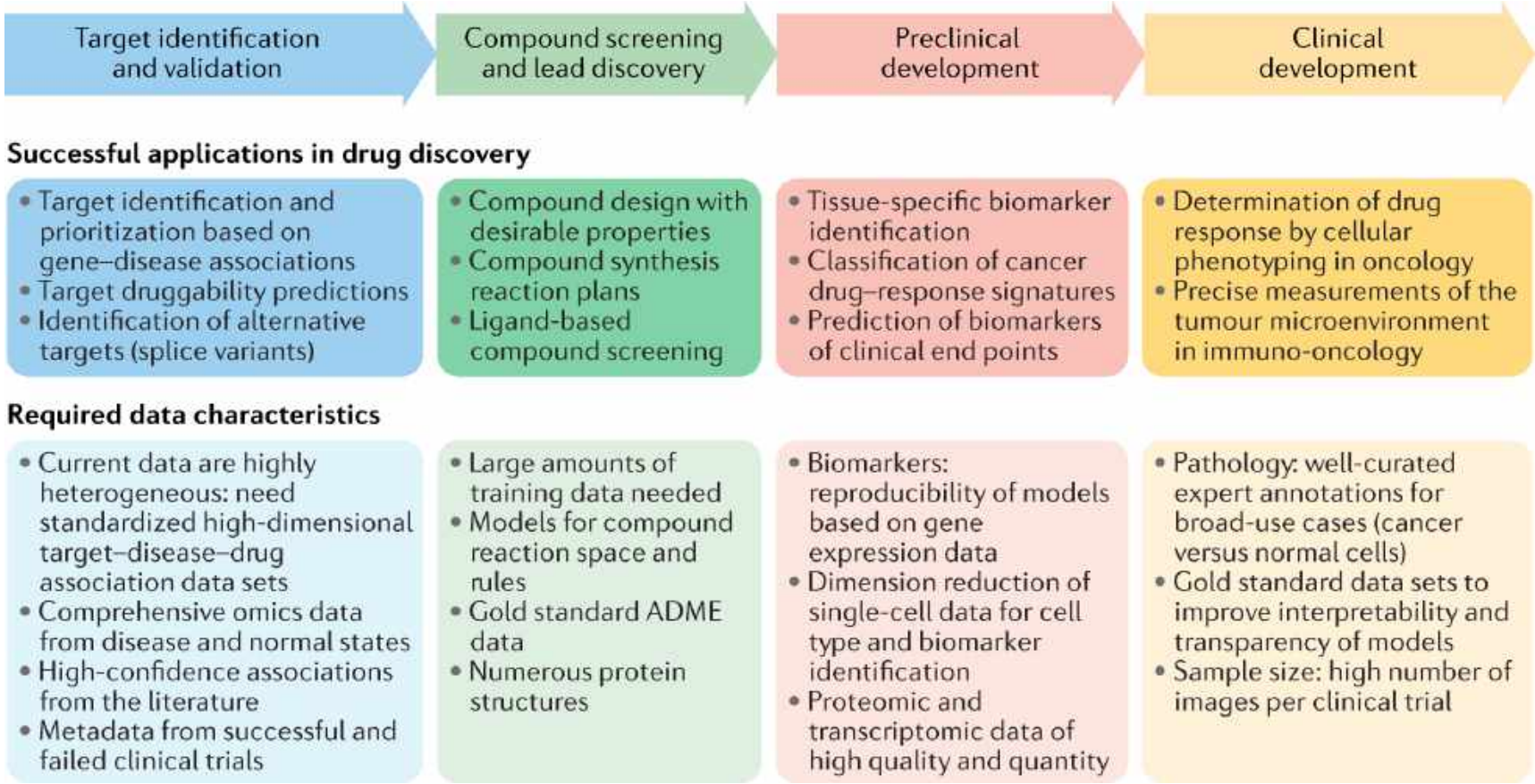


Sanchez-Lengeling et al., Science 361, 360–365 (2018)



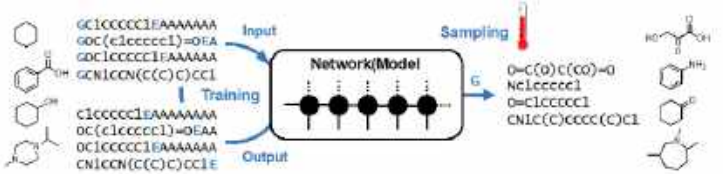
Ren *et al*. Sci Adv. (2017) eaaq1566

Machine Learning In Drug Discovery

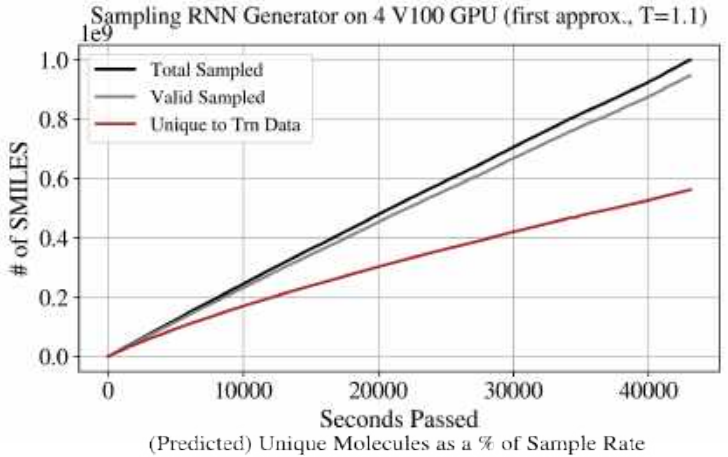


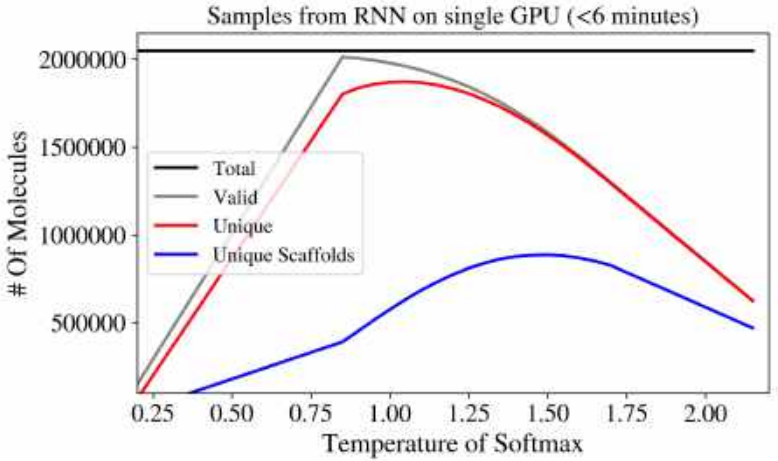
Integrating ML:
Pipelines that are ML designed.

RNN SMILES Modeling

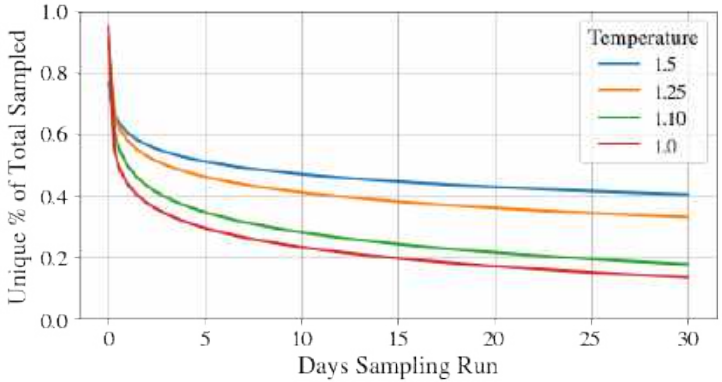


Gupta, Anvita, et al. "Generative recurrent networks for de novo drug design." Molecular informatics 37.1-2 (2018): 1700111.



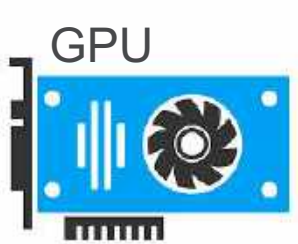


(Predicted) Unique Molecules as a % of Sample Rate

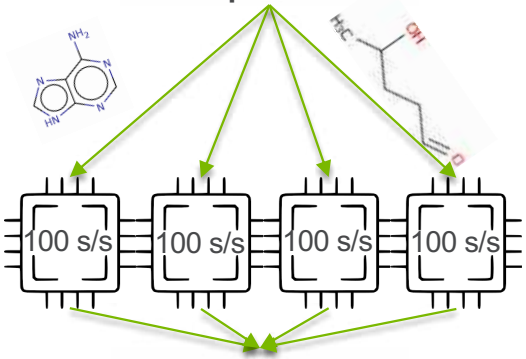


RNNs create a new paradigm for streaming development

- Rather than randomly drawing compounds from a library….
- What if we imagine a stream of compounds, all with some good properties, and then we want to work on those?



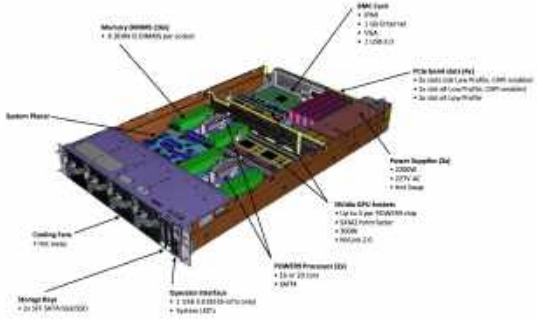
10,000 per second



Super fast, modern generative algorithms

Single threaded algorithms for CPU post-processing

Even slower simulations



IBM AC922, 6 GPU node. Balanced Heavily towards GPU, not CPU

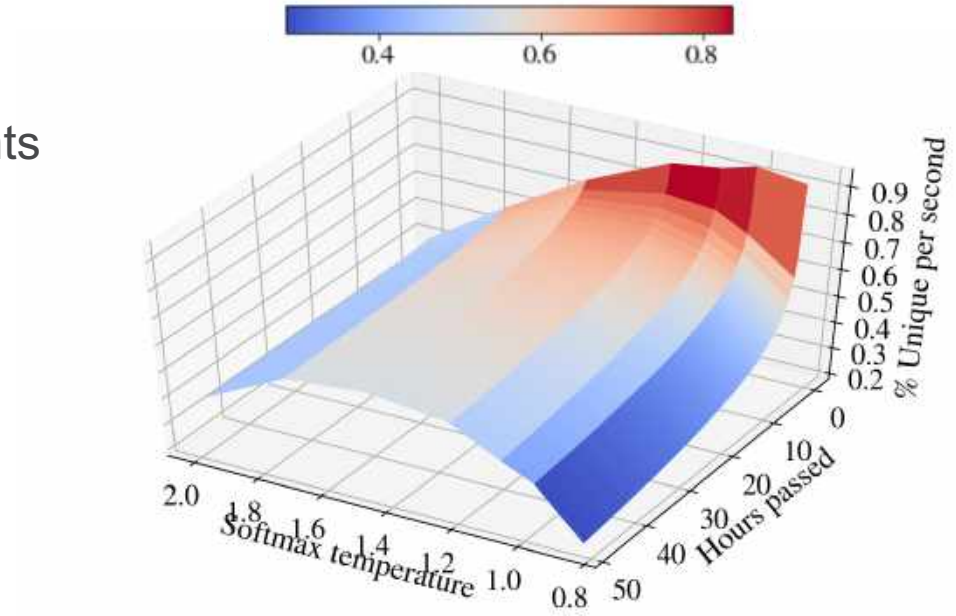
5000 Seconds per smiles

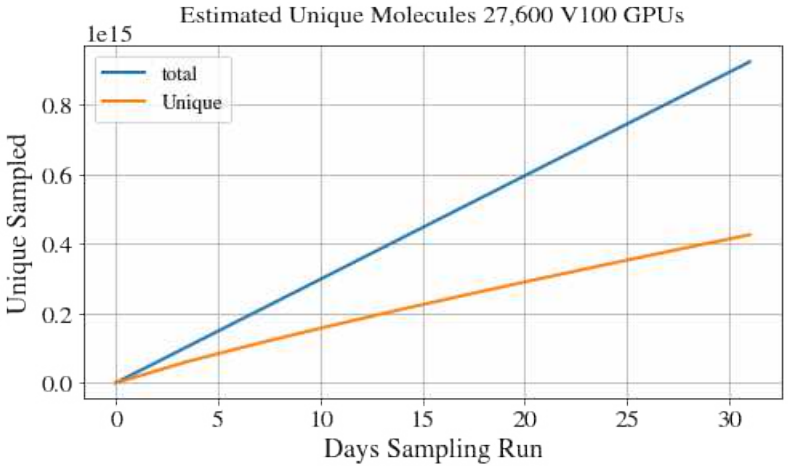
1 SMILE per second

In order to keep GPUs and CPUs hot, unique stream of molecules needs to stay constant

Top 0.001%

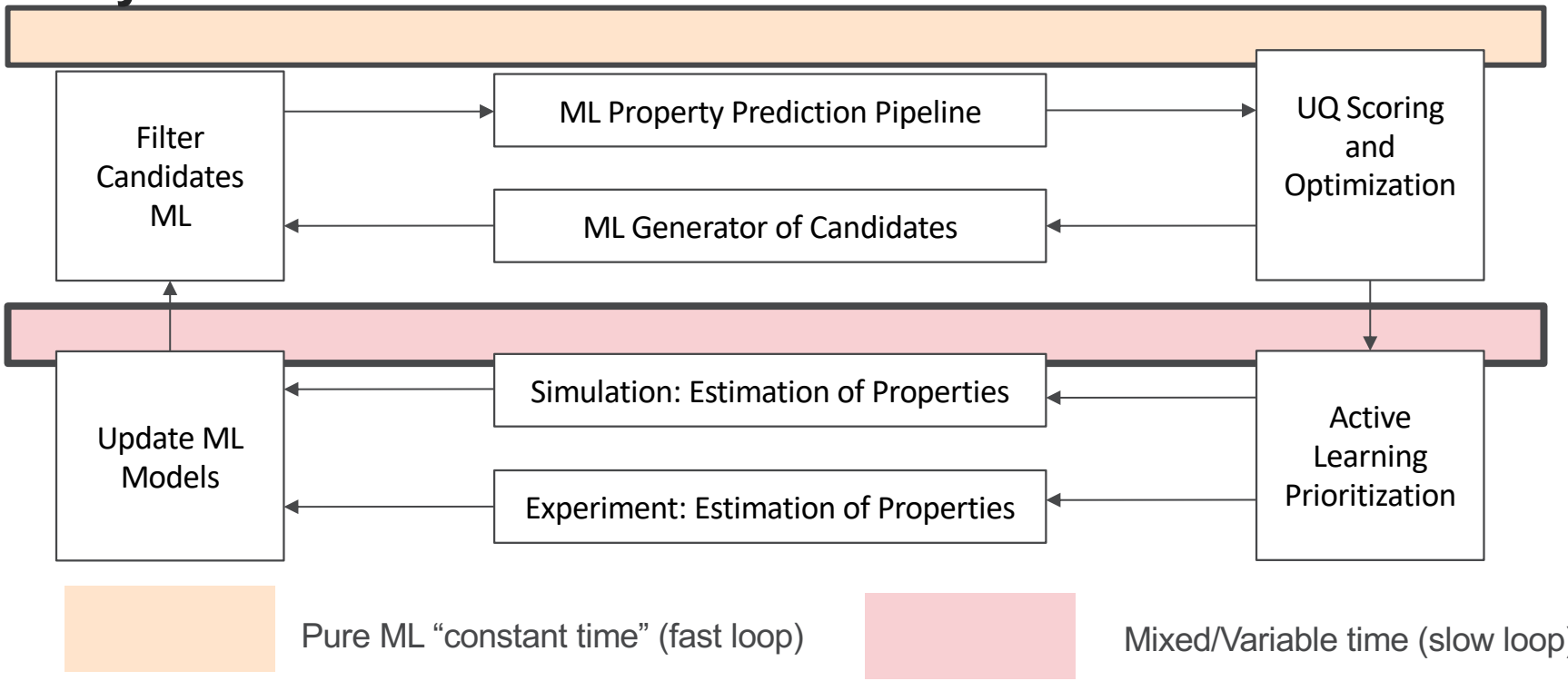
Database, Experiments





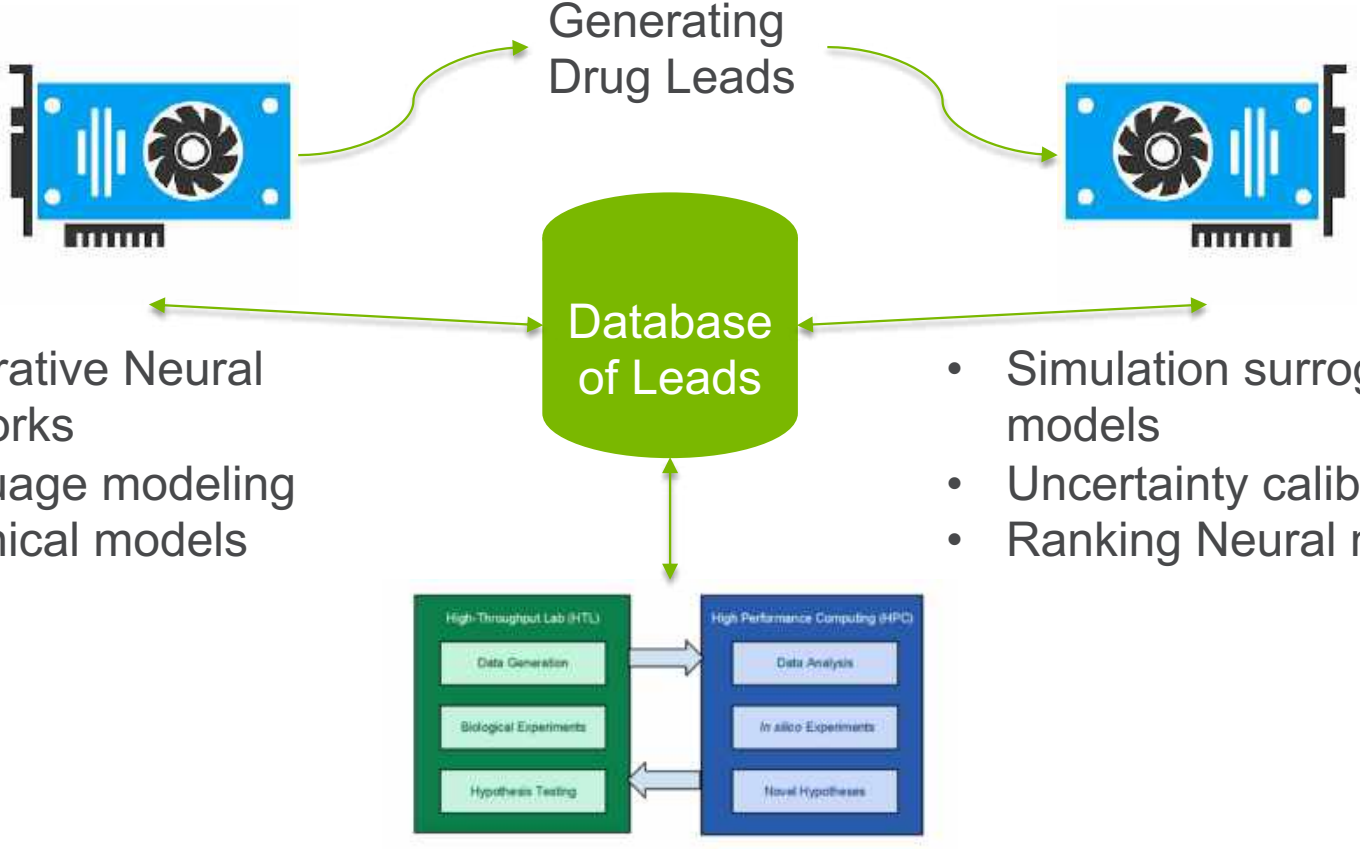
$$p(s)_i = \frac{e^{-\beta s_i}}{\sum_{j=0}^{K} e^{-\beta s_j}}$$

Layered workflow



HIGH THROUGHPUT SCREENING

Drug discovery



- Generative Neural Networks
- Language modeling
- Graphical models

- Simulation surrogate models
- Uncertainty calibrated
- Ranking Neural networks

But there is a sense we are just creeping

Ahead with baby steps in improvement

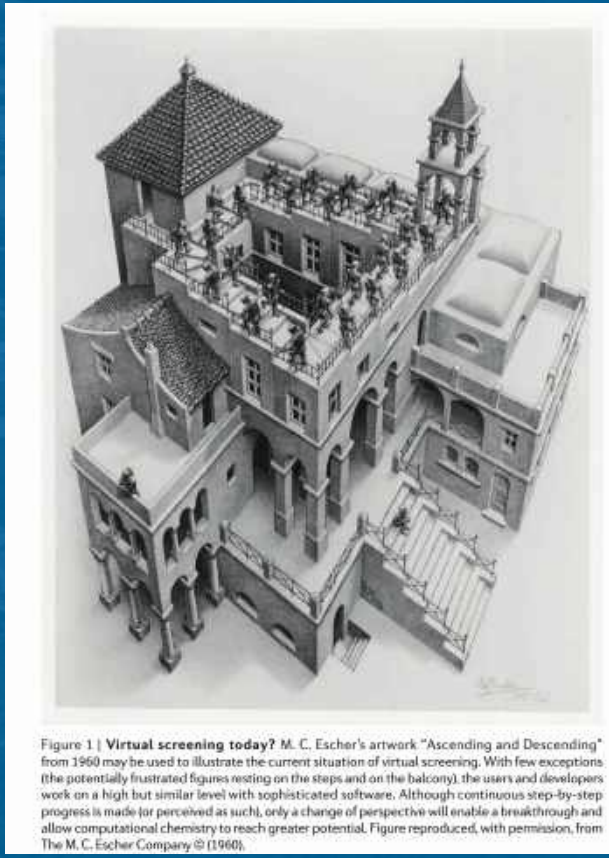


Figure 1 | **Virtual screening today?** M. C. Escher's artwork "Ascending and Descending" from 1960 may be used to illustrate the current situation of virtual screening. With few exceptions (the potentially frustrated figures resting on the steps and on the balcony), the users and developers work on a high but similar level with sophisticated software. Although continuous step-by-step progress is made (or perceived as such), only a change of perspective will enable a breakthrough and allow computational chemistry to reach greater potential. Figure reproduced, with permission, from The M. C. Escher Company © (1960).

Argonne NATIONAL LABORATORY

RL-Dock

Reinforcement Learning Docking and Simulation

- As a newcomer, the process seems lengthy…
  - Come up leads
  - Find targets and crystal structures
  - Look at conformers
  - Hours of compute simulating compound
  - Have to start over for new optimization or when toxicity testing comes

> Dynamic descriptions of molecules will have to replace our predominantly static view of both targets and ligands.

Gisbert Schneider

Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Swiss Federal Institute of Technology (ETH)

Designing a drug inside a flexible pocket in seconds?

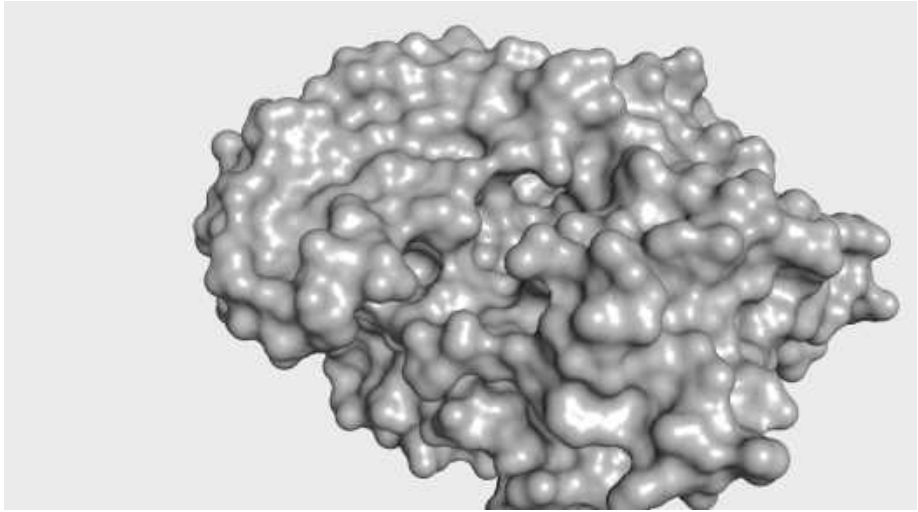
RESEARCH ARTICLE COMPUTATIONAL BIOLOGY

Deep reinforcement learning for de novo drug design

Mariya Popova[1,2,3], Olexandr Isayev[1,*] and Alexander Tropsha[1,*]

+ See all authors and affiliations

Science Advances 25 Jul 2018:
Vol. 4, no. 7, eaap7885
DOI: 10.1126/sciadv.aap7885

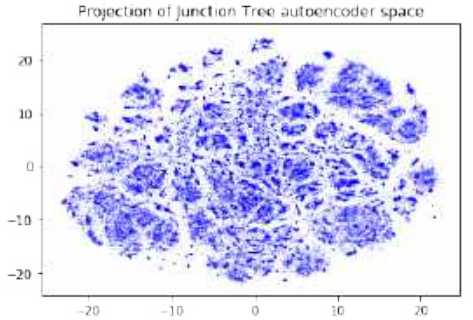


Connecting HPC and AI

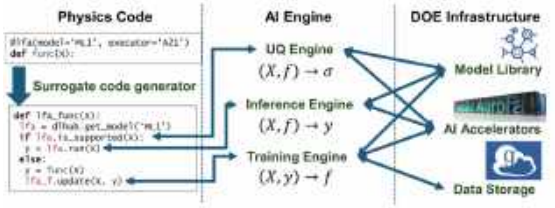
In addition to partnerships in AI applications, there are considerable opportunities in foundational methods development, software and software infrastructure for AI workflows and advanced hardware architectures for AI, below we highlight some ideas in the HPC + AI space

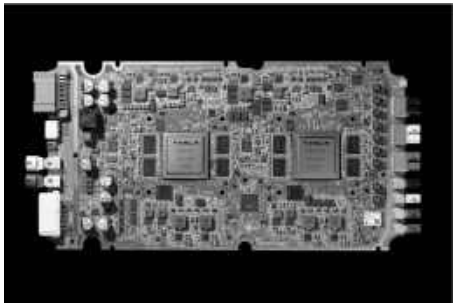
- Steering of simulations
- Embedding simulation into ML methods
- Customized computational kernels
- Tuning applications parameters
- Generative models to compare with simulation
- Student (AI) Teacher (Sim) models ⟹learned functions
- Guided search through parameter spaces
- Hybrid architectures HPC + Neuromorphic
- Many, many more

Generative Models



Learned Function Accelerators





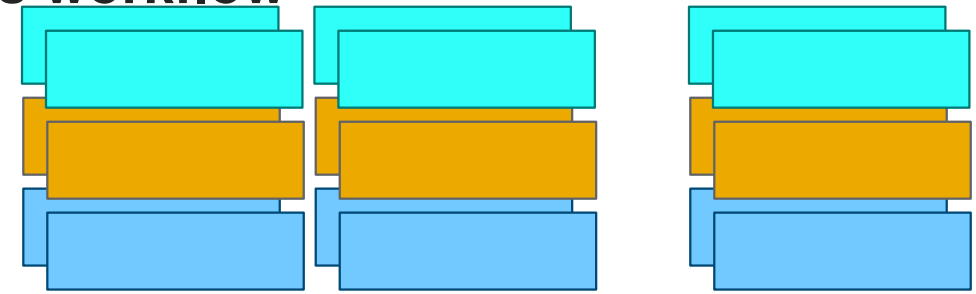
AI Accelerators

2

Main workflow → Epitope analysis workflow

Identify interacting regions and potentially mutated regions [COVID-19, SARS, MERS]

Evolution group provides inputs



Small molecule vs. mutating regions across coronaviruses

Apo-state (ligand-free state) simulations for [COVID-19, SARS, MERS]

Adaptive MD to determine protein-ligand specificity

DeepDriveMD

Exploit similarities in ligand binding to identify novel states

Runs on OLCF Summit

Target novel states for docking and refinement

0

Q&A

To pose a question, you can write your question in the “Questions” tab

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 675451

Thank you for participating!

…don't forget to fill in our feedback questionnaire…

Visit the CompBioMed website (www.compbiomed.eu/training)
for a full recording of this and other webinars,
to download the slides
and to keep updated on our upcoming trainings

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675451

The webinar series is run in collaboration with: VPH Institute Building the Virtual Physiological Human

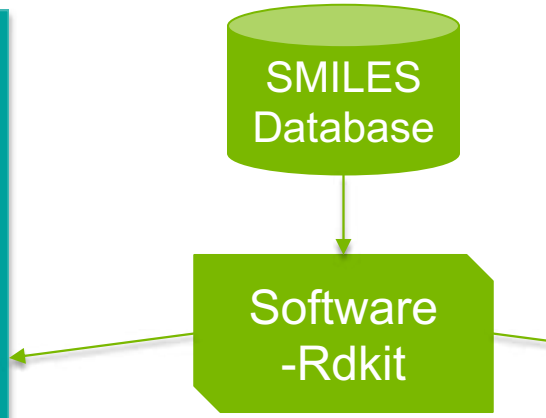
QUESTIONS?

aclyde@anl.gov

Molecular modalities

What *is* a "molecule" in the sense we're after?

- 2D Graphs
- 3D Coordinates
- 2D Images
- 3D Images (voxels)
- SMILES
  - Canonical
  - Kekule
- SELFIES (L2 Chomsky)
- Surface
- Conformer Sets

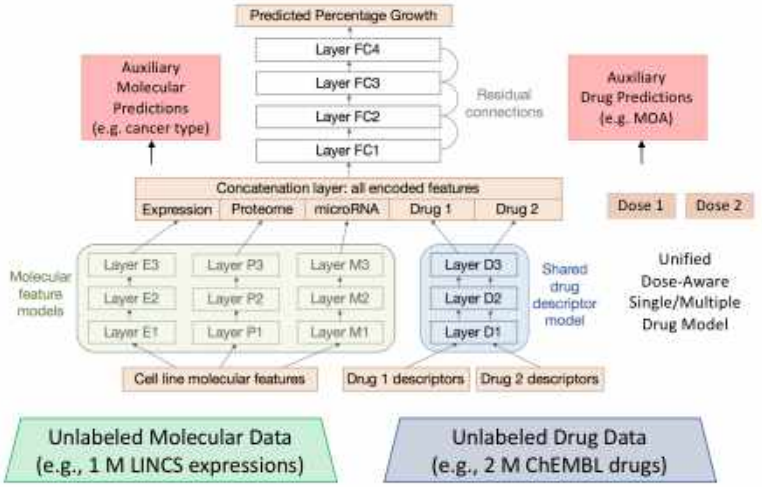


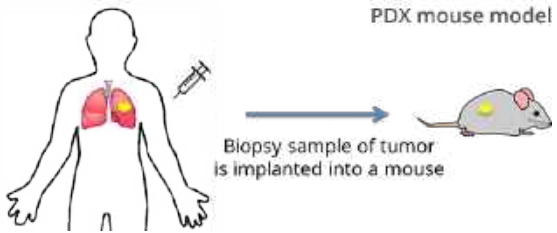
how do we chose?

- CFG models
- Docked models
- Chemical descriptors
- Fingerprints
- 3D waveform
- 3D density
- MACCS
- …many more in literature

Deep Learning Enabled Precision Medicine for Cancer
Rick Stevens (ANL PI), with NCI, LANL, LLNL, ORNL

- Opportunity
  - Increasing biological data require supercomputer-powered deep learning models in challenging cancer problems
- Argonne assets
  - Scalable machine learning
  - Expertise in computational biology and cheminformatics
- Strategy
  - Develop an exascale deep learning framework (CANDLE)
  - Build drug response models combining all available information on cancer types and drugs
- ML/DL methods applied
  - ResNets, ConvNets, MC-dropout, population based training
  - Semi-supervised deep learning, generative models
- Results achieved
  - Dose response model for drug pairs with 94% validation R^2
  - Large-scale inference runs with uncertainty quantification
  - Hyperparameter optimization tools
  - Drug candidates for patient-derived xenograft experiments

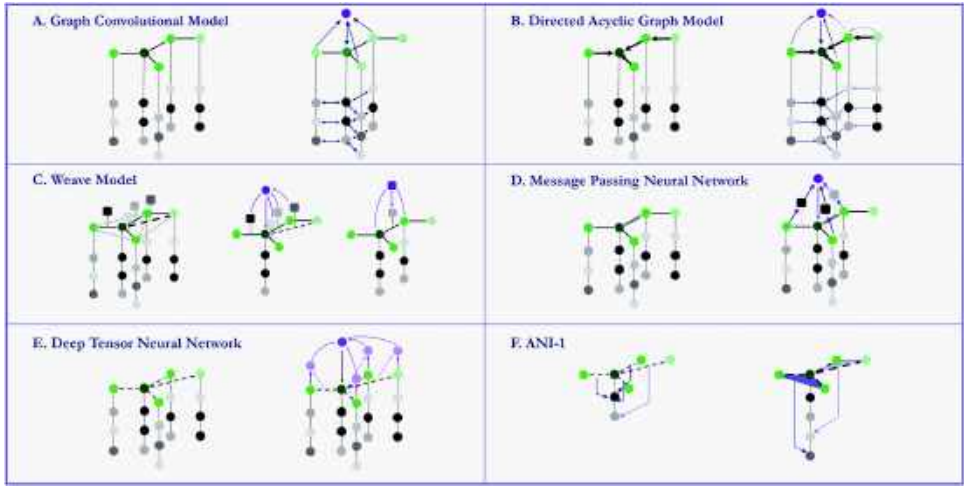


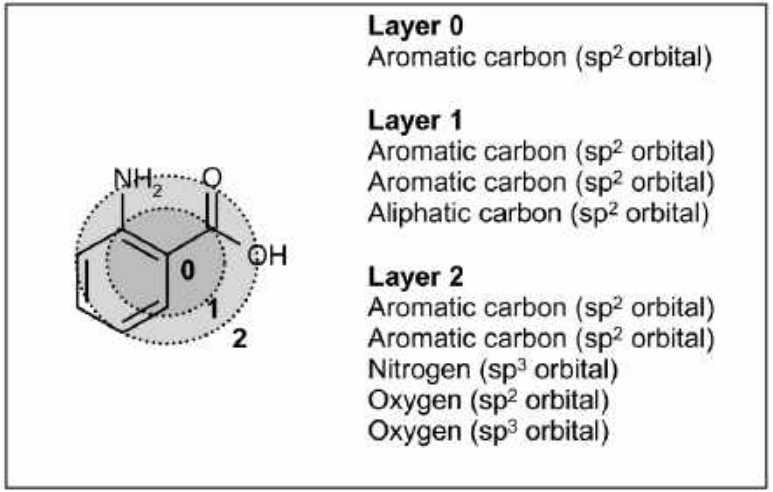


Graph-based networks

Provably more powerful than fingerprints

- Graph convolution is fundamentally more powerful than fingerprints





Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks?. *arXiv preprint arXiv:1810.00826*.

DFT (density functional theory) simulator

$\sim 10^3$ seconds

Toxic?

Bind with a given protein?

Quantum properties: $E, \omega_0, \ldots$

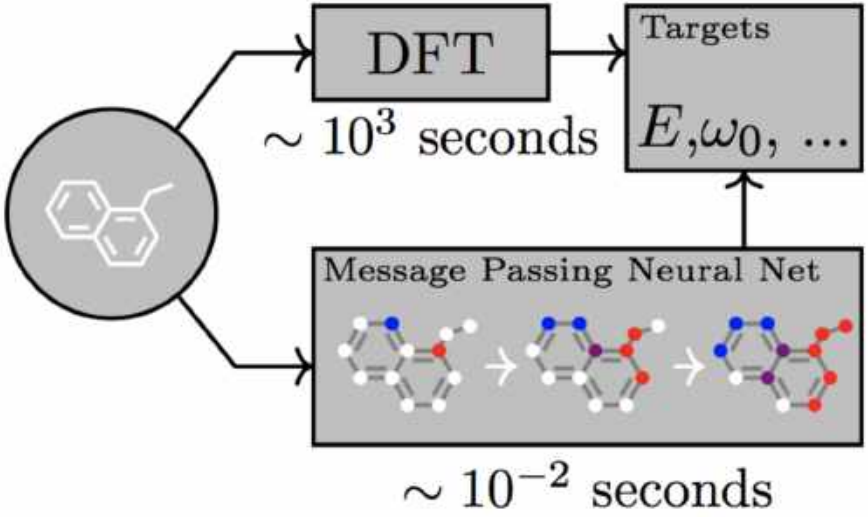
$$H\Psi = E\Psi = \sum_{i=1}^{N} \left(\frac{-\hbar^2}{2m} \nabla_i^2 \Psi - Ze^2 \sum_{\mathbf{R}} \frac{1}{|\mathbf{r}_i - \mathbf{R}|} \Psi \right) + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \Psi$$

- The red term describes correlation and is very difficult to account for
- The Hamiltonian can be generalised:

$$H = T + V + U$$

where U is the mutual interaction energy of the electrons and $V = \sum_{i=1}^{N} v(\mathbf{r}_i)$, the interaction with an arbitary external field.

Pierre Hohenberg and Walter Kohn 1964 — density functional theory



Neural Message Passing for Quantum Chemistry

Justin Gilmer [1] Samuel S. Schoenholz [1] Patrick F. Riley [2] Oriol Vinyals [3] George E. Dahl [1]

April 2017, >650 citations

ARGONNE NATIONAL LAB

Advancing basic science and engineering to benefit the U.S.

Argonne serves the U.S. as a science and energy laboratory distinguished by the breadth of our R&D capabilities and powerful suite of experimental and computational facilities

Argonne's mission is to **advance basic science and engineering** to benefit the U.S. economy and national security, it is not commercialization but strictly R&D

Argonne researchers **conceive and develop new technologies, and transfer those technologies** to the partners that can deliver the greatest positive impact to the nation and the world

Argonne's interests in commercialization do not conflict with industry, **we seek complementary business relationships**

Argonne NATIONAL LABORATORY

3,200 employees
1,300 scientists and engineers
260 postdoctoral researchers
7,920 facility users
$830M FY18 operating budget

ARGONNE LEADERSHIP COMPUTING FACILITY

CELL ANALYSIS, MODELING AND PROTOTYPING FACILITY

MATERIALS ENGINEERING RESEARCH FACILITY

- High-performance Computing
- Computational Science
- Artificial intelligence
- Urban and building technologies
- Resiliency / Cyber Security
- Energy Storage

ADVANCED PHOTON SOURCE USER FACILITY

ELECTROCHEMICAL ANALYSIS AND DIAGNOSTICS LABORATORY

- Connected and autonomous vehicles and e-mobility
- Electric vehicles
- Engines, fuels, emissions
- Smart manufacturing

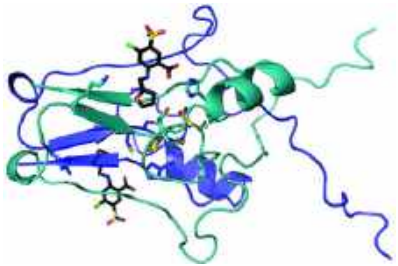
CENTER FOR NANOSCALE MATERIALS

- Materials characterization
- Nanomaterials

ADVANCED PHOTON SOURCE USER FACILITY

APS has supported pharma research since its opening in 2001

The Advanced Photon Source (APS) at Argonne provides ultra-bright, high-energy storage ring-generated x-ray beams for research in almost all scientific disciplines. It is the brightest x-ray source operating in the US today.



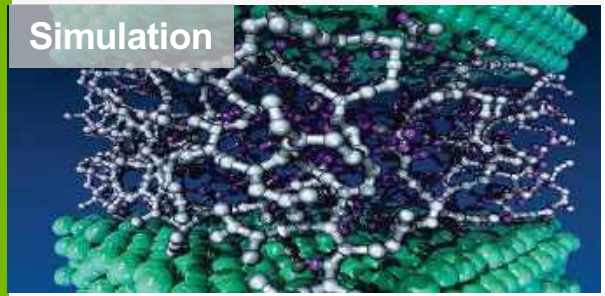
LRL-CAT: Lilly Research Laboratories operated dedicated beam-line

- Express Crystallography: a full-service mail-in program in protein crystallography to industrial, government and academic users of the APS

IMCA-CAT: Industrial Macromolecular Crystallography Association

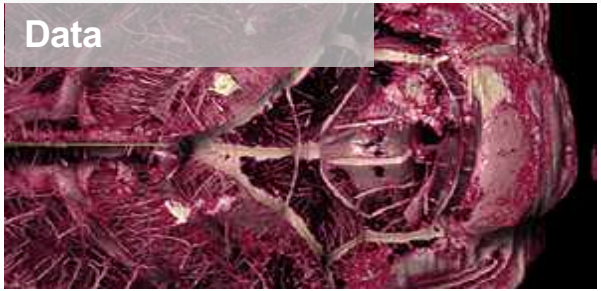
- Established in 1990, IMCA is committed to the use of macromolecular crystallography as a tool in drug discovery and product development.
- Managed througha contract with Hauptman-Woodward Medical Research Institute
- Member companies: AbbVie, Bristol-Myers Squibb, Merck, Novartis, and Pfizer
- Non-member companies are invited to collect proprietary data at the beamline via the IMCA-CAT subscription program

EXPANDING LEADERSHIP COMPUTING REACH



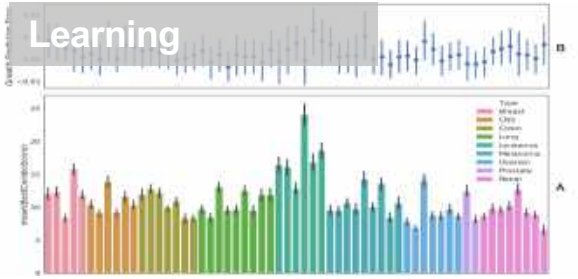
Reactive Mesoscale Simulations of Tribological Interfaces

PI: S. Sankaranarayanan, ANL
Insight to the complex processes that make oils, coatings, electrodes, and other electrochemical interfaces effective. Using Mira, this team discovered a self-healing, anti-wear coating that drastically reduces friction. Their findings are being used to virtually test other potential self-regenerating catalysts.



Large-Scale Computing on the Connectomes of the Brain

PI: D. Gursoy, ANL
3D reconstructions of high-resolution imaging will provide a clearer understanding of how even the smallest changes to the brain play a role in the onset and evolution of neurological diseases, such as Alzheimer's and autism, and perhaps lead to improved treatments or even a cure.



CANcer Distributed Learning Environment (CANDLE)

PI: R. Stevens, ANL
CANDLE is tackling the hardest deep learning problems in cancer research. Its first architecture release for large-scale model hyperparameter exploration uses representative problems--coded as deep learning problems--at the core of the predictive oncology challenge. Future data parallelism work will allow the training of a single model across several nodes.

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

ALCF DATA SCIENCE PROGRAM

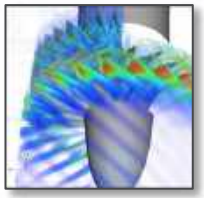
ECP
EXASCALE COMPUTING PROJECT

ECP applications target national problems in 6 strategic areas

National security

Stockpile stewardship

Next-generation electromagnetics simulation of hostile environment and virtual flight testing for hypersonic re-entry vehicles



Energy security

Turbine wind plant efficiency

High-efficiency, low-emission combustion engine and gas turbine design

Materials design for extreme environments of nuclear fission and fusion reactors

Design and commercialization of Small Modular Reactors

Subsurface use for carbon capture, petroleum extraction, waste disposal

Scale-up of clean fossil fuel combustion

Biofuel catalyst design

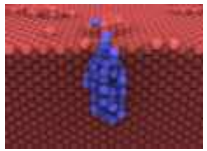
Economic security

Additive manufacturing of qualifiable metal parts

Reliable and efficient planning of the power grid

Seismic hazard risk assessment

Urban planning



Scientific discovery

Find, predict, and control materials and properties

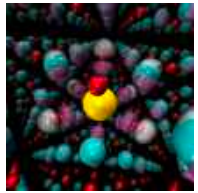
Cosmological probe of the standard model of particle physics

Validate fundamental laws of nature

Demystify origin of chemical elements

Light source-enabled analysis of protein and molecular structure and design

Whole-device model of magnetically confined fusion plasmas



Earth system

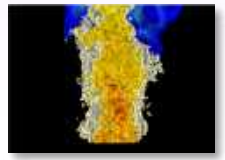
Accurate regional impact assessments in Earth system models

Stress-resistant crop analysis and catalytic conversion of biomass-derived alcohols

Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Health care

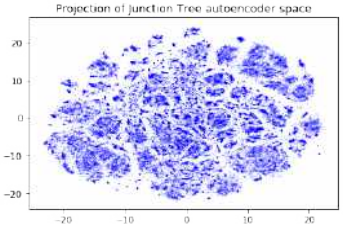
Accelerate and translate cancer research

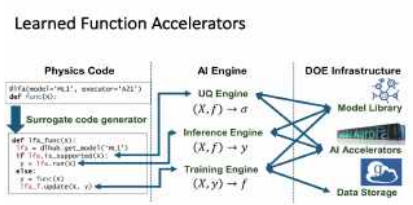


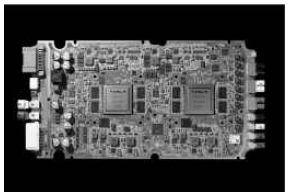
Wide ranging opportunities to Connect HPC and AI

- Steering of simulations
- Embedding simulation into ML methods
- Customized computational kernels
- Tuning applications parameters
- Generative models to compare with simulation
- Student (AI) Teacher (Sim) models ⟹learned functions
- Guided search through parameter spaces
- Hybrid architectures HPC + Neuromorphic
- Many others

Generative Models







AI Accelerators

Integrating HPC and AI with high-throughput experiments

To accelerate development of new materials, chemicals, proteins, pathways and organisms