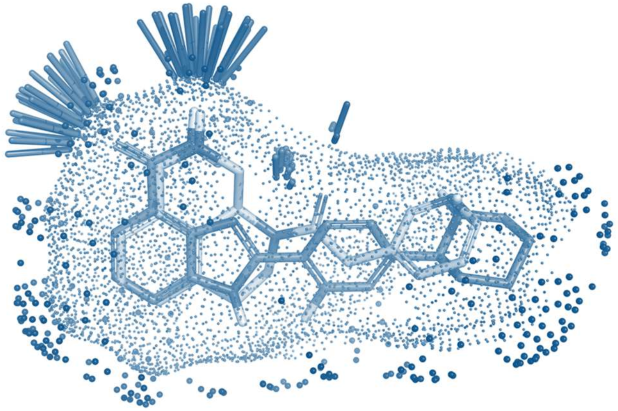
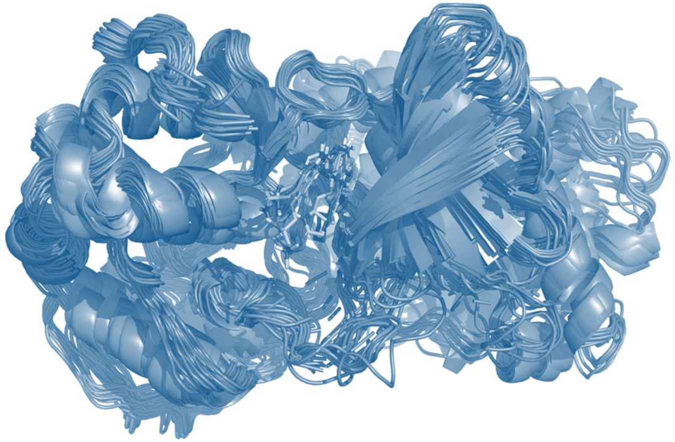




BioPharmics™

a division of optibrium™

Physical Parameter Estimation vs. Pure Machine-Learning for Drug Design

Ajay N. Jain, PhD, ajay@optibrium.com
Matt Segall, PhD, matt@optibrium.com
Himani Tandon, PhD, himani@optibrium.com
Ann E. Cleves, PhD, ann@optibrium.com

ACS Fall 2025, 8-18-2025, Washington DC

Machine Learning in CADD has Special Challenges

Therapeutic small molecules are only rarely experiments of nature!

CADD prediction challenges

- The things we want to predict are in the future (e.g. what a candidate molecule will do)
- They do not come from the same statistical population as the molecules/activity-data from which we can induce models
- This violates the central assumption of machine-learning:

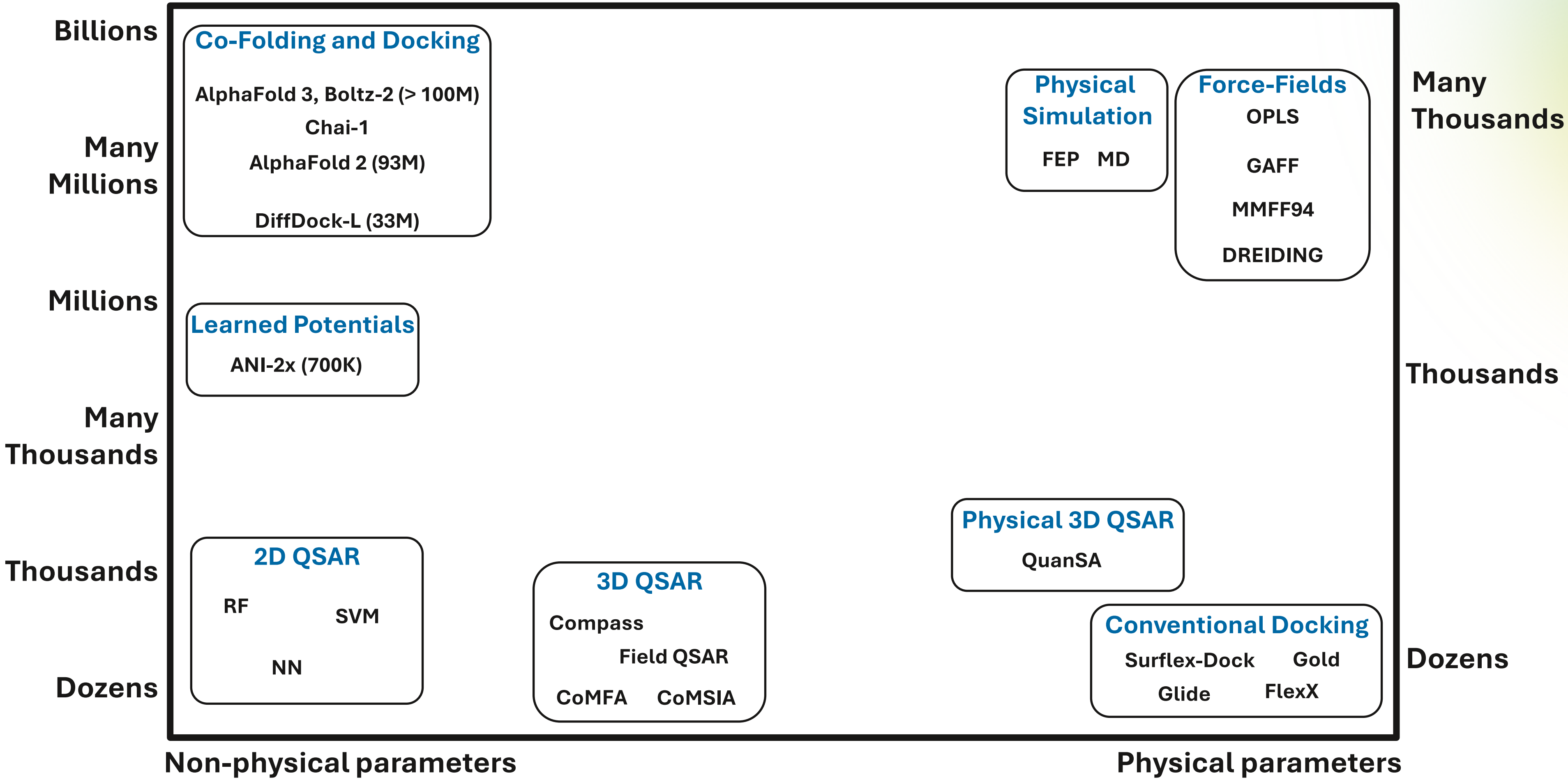
Predict on things that come from the *same population* as things used for training a model

Pure ML vs. Physical Parameter Estimation

- Pure machine learning
  - A numerical input representation may be grounded in physically relevant features for a particular domain
  - But the parameters to be estimated are inscrutable
  - Subject to the central ML assumption
- Physical parameter estimation
  - Begins from a model that mirrors physical reality
    - At the quantum level, we know the “truth” about atoms and molecules
    - We have developed extremely good approximations (e.g. DFT)
    - We have good grasp of non-covalent binding based on thermodynamics
  - Each parameter is directly related to a physical quantity
  - With physical realism, we might be able to make predictions on a causal basis: ***does not require*** population assumptions

Two Dimensions: Physicality vs. Number of Parameters

Parameter counts are of a different order with the newest Pure-ML models



Billions

Many Millions

Millions

Many Thousands

Thousands

Dozens

Co-Folding and Docking

AlphaFold 3, Boltz-2 (> 100M)

Chai-1

AlphaFold 2 (93M)

DiffDock-L (33M)

Learned Potentials

ANI-2x (700K)

2D QSAR

RF SVM NN

3D QSAR

Compass

Field QSAR

CoMFA CoMSIA

Physical Simulation

FEP MD

Force-Fields

OPLS

GAFF

MMFF94

DREIDING

Physical 3D QSAR

QuanSA

Conventional Docking

Surflex-Dock Gold

Glide FlexX

Many Thousands

Thousands

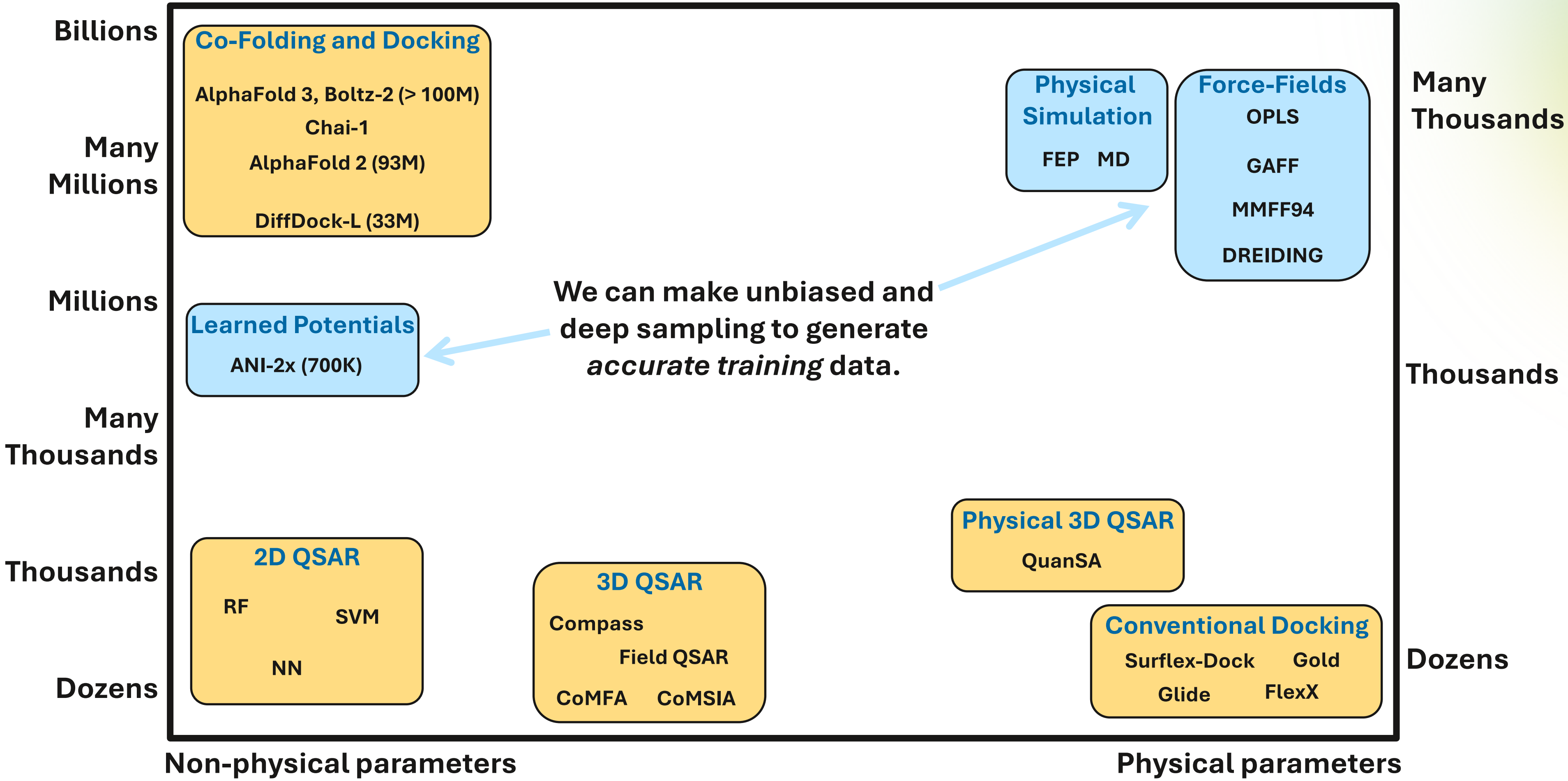
Dozens

Non-physical parameters

Physical parameters

Actually More Than Two Dimensions

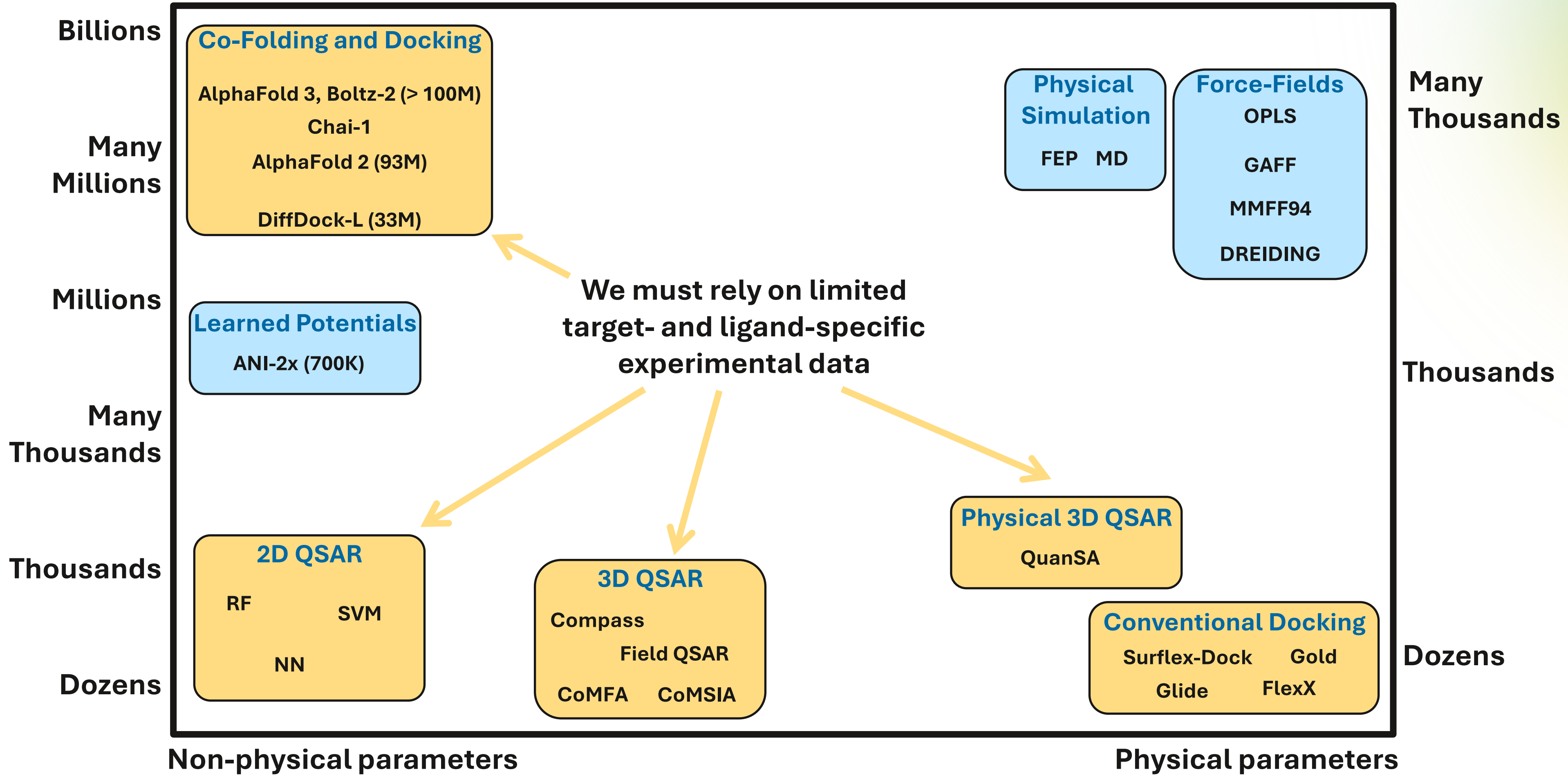
Dependency on experimental data is another dimension



We can make unbiased and deep sampling to generate *accurate training* data.

Actually More Than Two Dimensions

Dependency on experimental data is another dimension



Billions

Many Millions

Millions

Many Thousands

Thousands

Dozens

Co-Folding and Docking

AlphaFold 3, Boltz-2 (> 100M)

Chai-1

AlphaFold 2 (93M)

DiffDock-L (33M)

Learned Potentials

ANI-2x (700K)

Physical Simulation

FEP MD

Force-Fields

OPLS

GAFF

MMFF94

DREIDING

We must rely on limited target- and ligand-specific experimental data

2D QSAR

RF

SVM

NN

3D QSAR

Compass

Field QSAR

CoMFA CoMSIA

Physical 3D QSAR

QuanSA

Conventional Docking

Surflex-Dock Gold

Glide FlexX

Many Thousands

Thousands

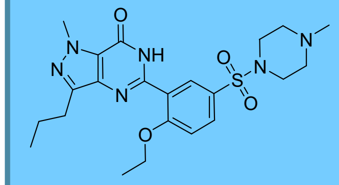
Dozens

Non-physical parameters

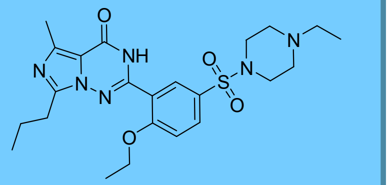
Physical parameters

Target choice and ligand structure reflect economics, fashion, and human design bias

Ligands for the same target change dramatically over time



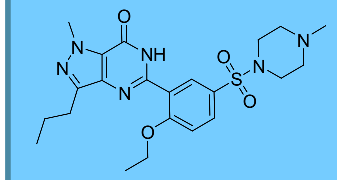
Sildenafil Citrate (Viagra)
Pfizer
Patented 1992



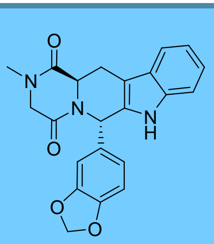
Vardenafil (Levitra)
Marketed by Bayer/GSK
Patented 2002

Viagra/Levitra is the norm

If future drugs against a target came from the same population as past ones, there would be no distributional difference in the blue and magenta curves.

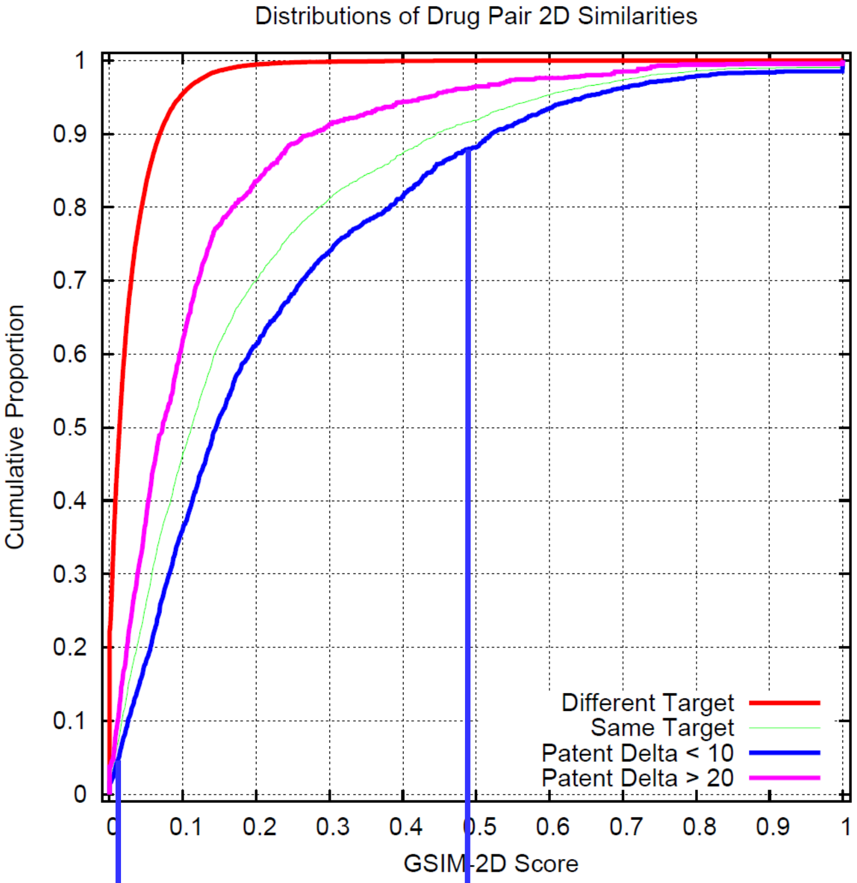


Sildenafil Citrate (Viagra)
Pfizer
Patented 1992



Tadalafil (Cialis)
Marketed by Lilly
Patented 1995
Markedly better PK.

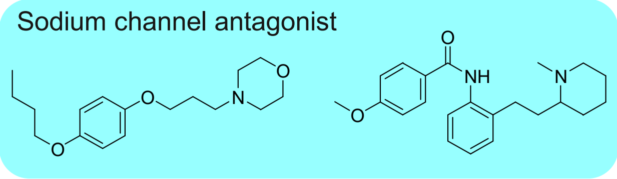
Viagra/Cialis is the outlier



Cleves, A.E., Jain, A.N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. J Comput Aided Mol Des 22, 147–159 (2008). https://doi.org/10.1007/s10822-007-9150-y

Ligand design reflects 2D thinking: A human inductive bias

The only difference between the **cyan** and **green** curves is that humans were thinking about the same target for the green pairs.

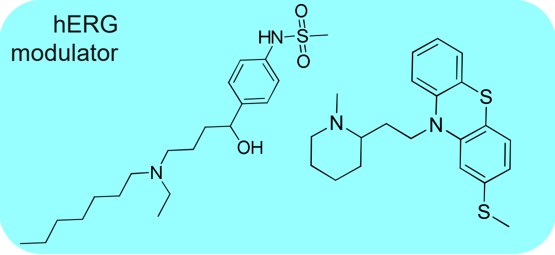


Pramoxine

Encainide

Potassium channel antagonist (cardiac)

Also hits sodium channel

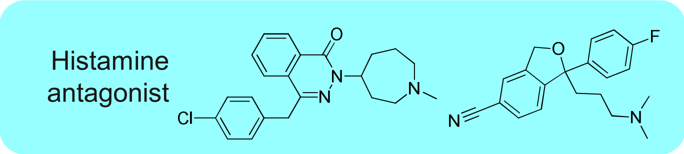


Ibutilide

Thioridazine

Dopamine antagonist

Also hits hERG

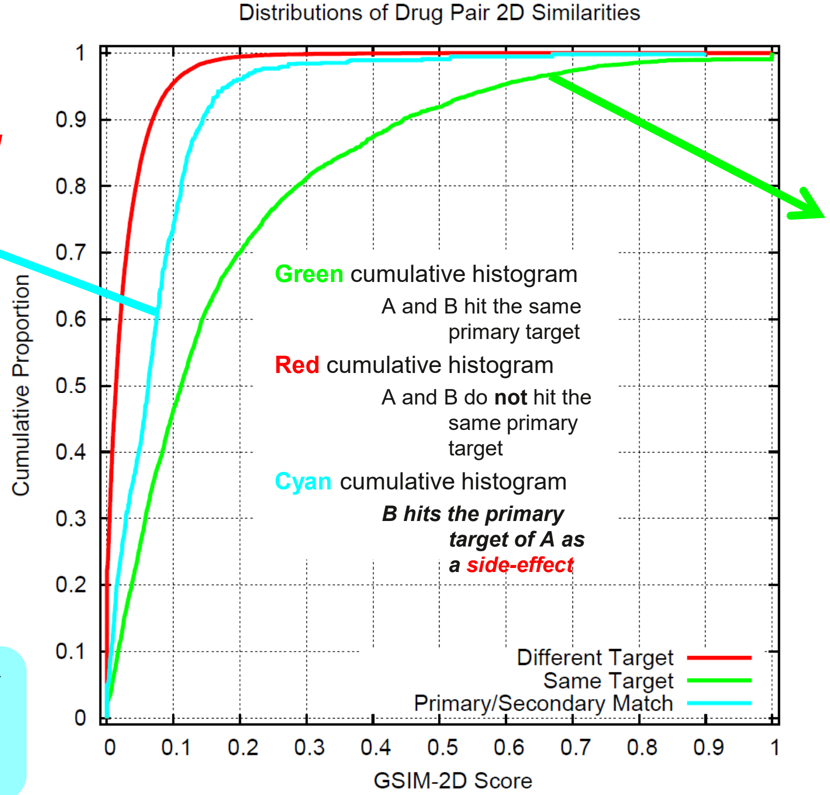


Azelastine

Citalopram

Serotonin reuptake inhibitor

Also hits histamine receptor



Green cumulative histogram

A and B hit the same primary target

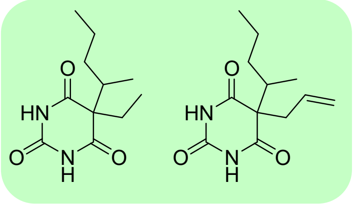
Red cumulative histogram

A and B do **not** hit the same primary target

Cyan cumulative histogram

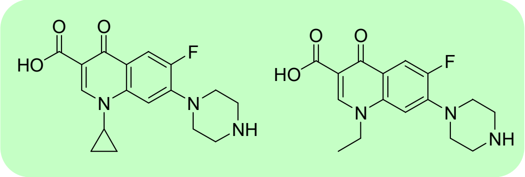
B hits the primary target of A as a side-effect

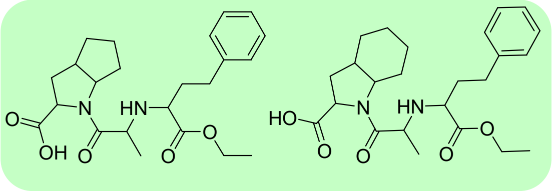
These 2D-influenced design examples are *hugely* overrepresented in our data sets!



GABA$_A$ barbiturate-site

DNA gyrase



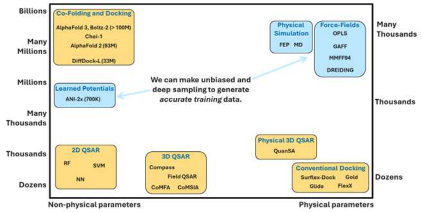


β Adrenergic

Cleves, A.E., Jain, A.N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. J Comput Aided Mol Des 22, 147–159 (2008). https://doi.org/10.1007/s10822-007-9150-y

Molecular Mechanics Potentials

Physical parameter estimation relies on a sensible model of molecules



Physical model

- Atoms and bonds, with assigned types
  - Atoms (1 atom)
  - Bonds (2 atoms)
  - Bond angles (3 atoms)
  - Torsions (4 atoms)
  - Non-bonded interactions (2 atoms)
- Relatively simple functions with internal parameters to estimate
- Many thousands of parameters

Among the most successful predictive modeling approaches

Many variations!

- AMBER (GAFF): https://doi.org/10.1021/acs.jpcb.5b00689
- MMFF94 https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3C490::AID-JCC1%3E3.0.CO;2-P
- OPLS3 https://doi.org/10.1021/acs.jctc.5b00864

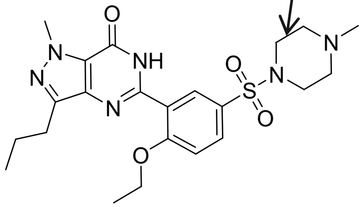
Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94*

THOMAS A. HALGREN
Department of Molecular Design and Diversity, Merck Research Laboratories, Rahway, New Jersey 07065

$$E_{MMFF} = \sum EB_{ij} + \sum EA_{ijk} + \sum EBA_{ijk} + \sum EOOP_{ijk;l} + \sum ET_{ijkl} + \sum EvdW_{ij} + \sum EQ_{ij}$$

$$EB_{ij} = 143.9325 \frac{\boxed{kb_{IJ}}}{2} \Delta r_{ij}^2 \times \left(1 + cs\,\Delta r_{ij} + 7/12 cs^2 \Delta r_{ij}^2\right)$$

$$r_{ij} - \boxed{r^0_{IJ}}$$



JCTC Journal of Chemical Theory and Computation Article
pubs.acs.org/JCTC

OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins

Edward Harder,[*,†] Wolfgang Damm,[†] Jon Maple,[†] Chuanjie Wu,[†] Mark Reboul,[†] Jin Yu Xiang,[†] Lingle Wang,[†] Dmitry Lupyan,[†] Markus K. Dahlgren,[†] Jennifer L. Knight,[†] Joseph W. Kaus,[†] David S. Cerutti,[†] Goran Krilov,[†] William L. Jorgensen,[§] Robert Abel,[†] and Richard A. Friesner[‡]

[†]Schrodinger, Inc., 120 West 45th Street, New York, New York 10036, United States
[‡]Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027, United States
[§]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

$$E = \sum_{i<j} [q_i q_j e^2/r_{ij} + 4\varepsilon_{ij}(\sigma_{ij}^{12}/r_{ij}^{12} - \sigma_{ij}^6/r_{ij}^6)]f_{ij} + \sum_{\text{bonds}} \boxed{K_r}(r - \boxed{r_{\text{eq}}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \left[\frac{V_1}{2}(1 + \cos\varphi) + \frac{V_2}{2}(1 - \cos 2\varphi) + \frac{V_3}{2}(1 + \cos 3\varphi) + \frac{V_4}{2}(1 - \cos 4\varphi)\right]$$

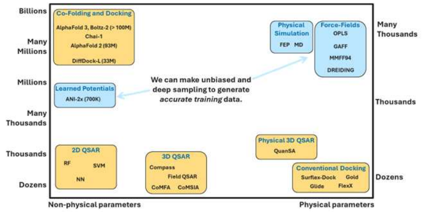
Table 1. Number of Unique Parameters for Valence Terms in the Respective Force Fields

parameter type	MMFF	OPLS_2005	OPLS2.1	OPLS3
stretches	456	1054	1181	1187
bends	2283	3997	14916	15236
torsions	520	1576	45472	48142

The parameters are estimated using both experimental and quantum mechanical data, the latter being carefully generated to cover the desired chemical space.

Pure-ML Energetic Potentials

Black-box parameter estimation relies on **MANY** training examples

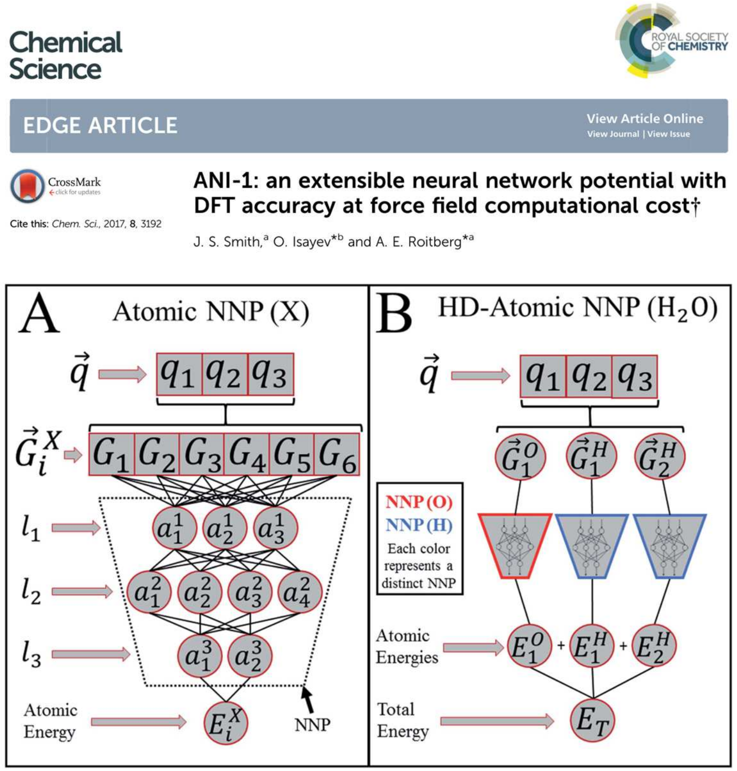


ANI-1

- Accurate NeurAl networK engINe for Molecular Energies (ANAKIN-ME)
- Parameterized for CHNO
- Computes an atomic-environment-vector
  - These probe specific regions of an individual atom's radial and angular chemical environment
- Must estimate > 100 thousand parameters
- Uses a huge amount of unbiased training data
  - Nearly 22,000,000 conformational energies
  - 57,000 molecules from the GDB-11 database, which exhaustively enumerates stable small molecules

ANI-2X

- Generalizes to seven elements: (H, C, N, O, F, Cl, S)
- Roughly 700,000 parameters
- Uses *active learning* to choose training exemplars (millions)



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

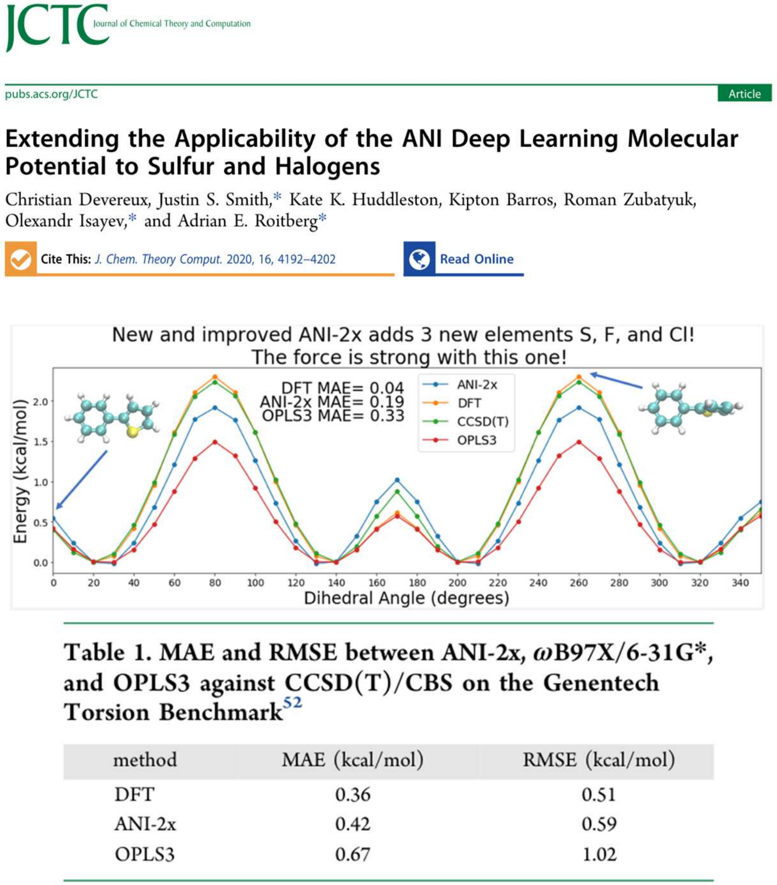
View Article Online
View Journal | View Issue

CrossMark

Cite this: Chem. Sci., 2017, 8, 3192

ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost†

J. S. Smith,[a] O. Isayev*[b] and A. E. Roitberg*[a]



JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC

Article

Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens

Christian Devereux, Justin S. Smith,* Kate K. Huddleston, Kipton Barros, Roman Zubatyuk, Olexandr Isayev,* and Adrian E. Roitberg*

Cite This: J. Chem. Theory Comput. 2020, 16, 4192−4202

Read Online

New and improved ANI-2x adds 3 new elements S, F, and Cl!
The force is strong with this one!

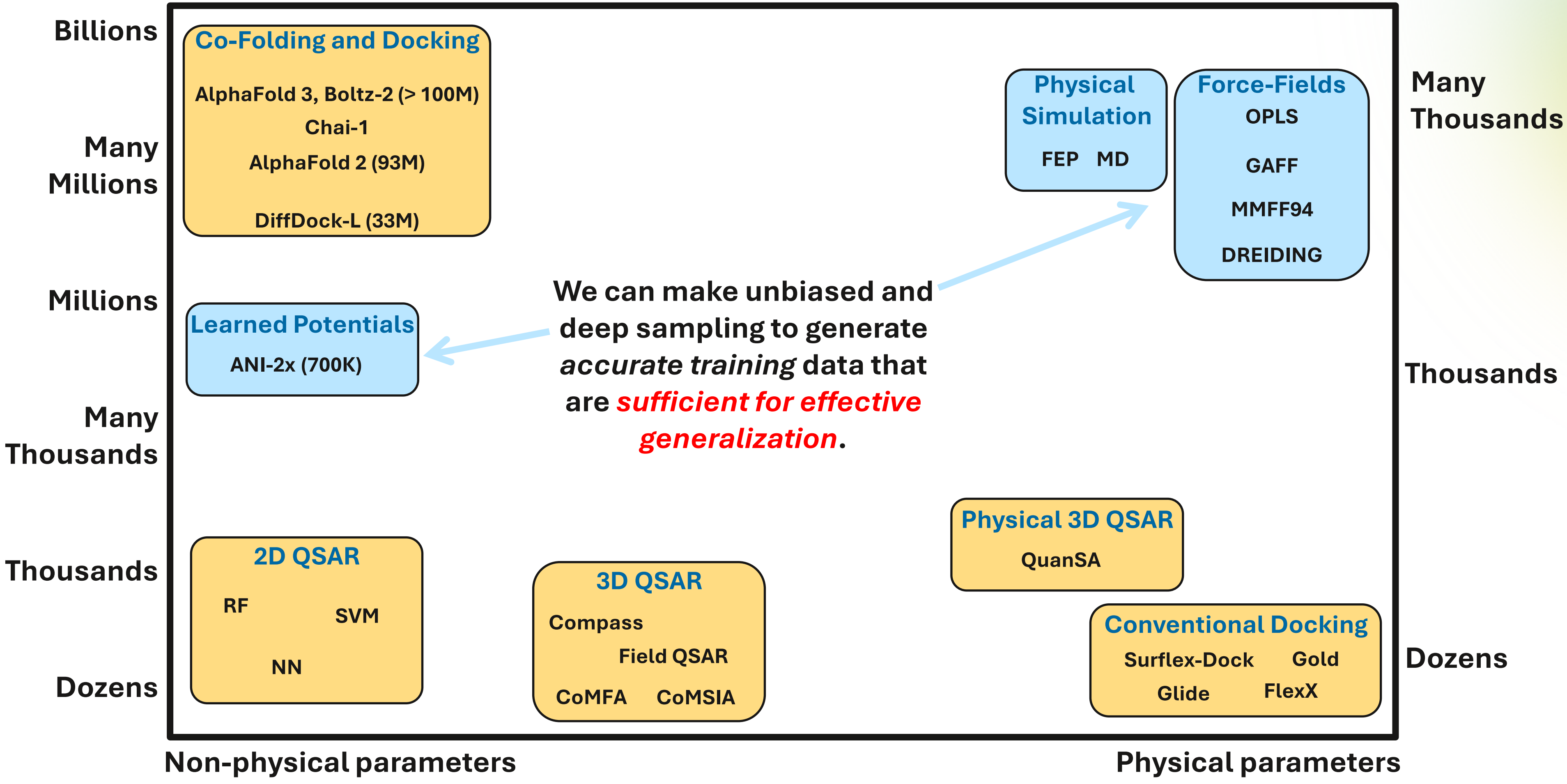
Table 1. MAE and RMSE between ANI-2x, ωB97X/6-31G*, and OPLS3 against CCSD(T)/CBS on the Genentech Torsion Benchmark[52]

method	MAE (kcal/mol)	RMSE (kcal/mol)
DFT	0.36	0.51
ANI-2x	0.42	0.59
OPLS3	0.67	1.02

The parameters are estimated using massive and unbiased data sets of DFT-based conformational energies.

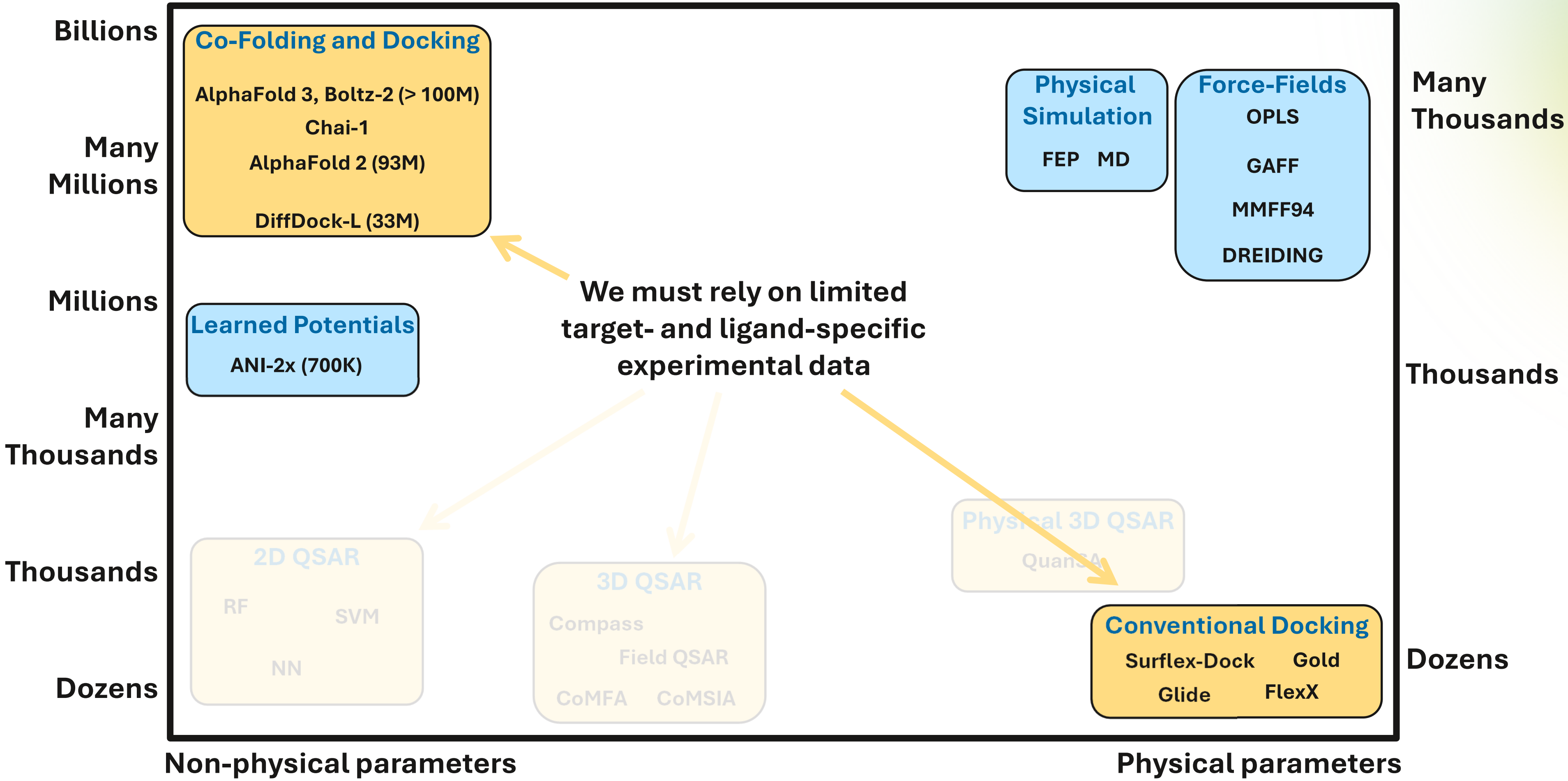
Huge, accurate, and unbiased training sets

Pure ML learned potentials and physically parameterized force-fields are successful and beneficial



We can make unbiased and deep sampling to generate *accurate training* data that are ***sufficient for effective generalization***.

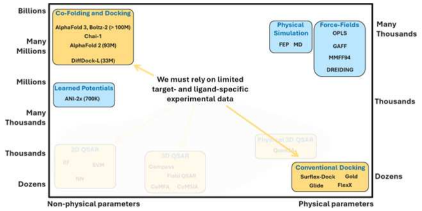
What happens when we must rely on experimental data?



We must rely on limited target- and ligand-specific experimental data

Co-Folding: Pure ML strongly affected by near-neighbor effects

Škrinjar, Eberhardt, Durairaj, Schwede 2025: AlphaFold3, Chai-1, Protenix, and Boltz-1



Benchmark

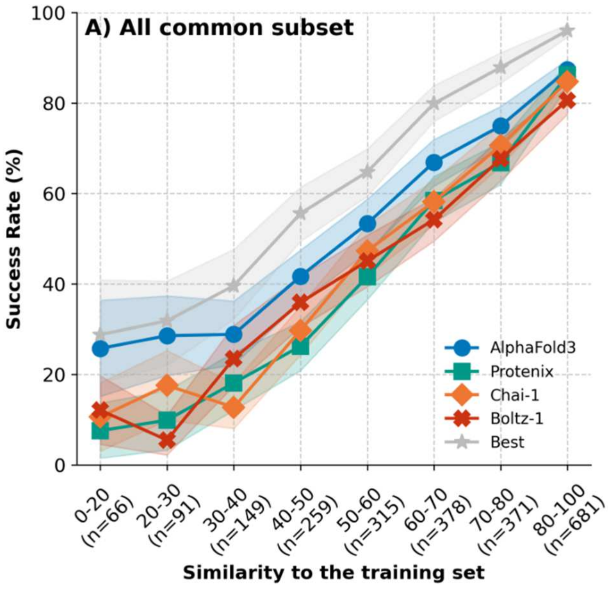
- 2600 protein/ligand structures post 9-30-2021
- The date cutoff was *after* training data for co-folding methods

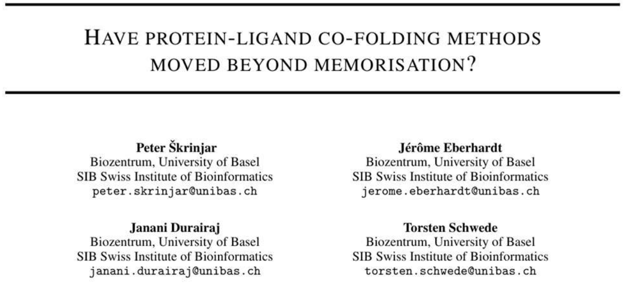
Pure ML

- AlphaFold3, Chai-1, Protenix, and Boltz-1
- Number of parameters: **Millions**
- Number of training exemplars: Tens of thousands

Observations echoed in multiple papers

- Matthew R. Masters, Amr H. Mahmoud, Markus A. Lill (2024) https://doi.org/10.1101/2024.06.03.597219
- Ajay N. Jain, Ann E. Cleves, W. Patrick Walters (2024) https://doi.org/10.48550/arXiv.2412.02889
- Martin Buttenschoen, Garrett M. Morris, Charlotte M. Deane (2023) https://doi.org/10.48550/arXiv.2308.05777





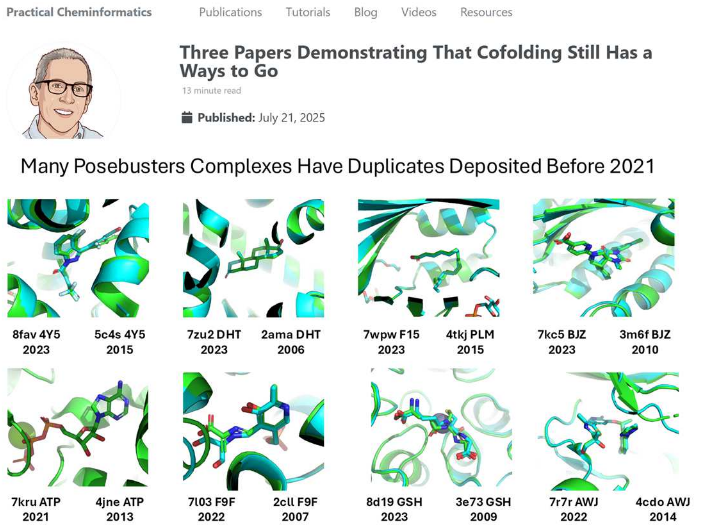
HAVE PROTEIN-LIGAND CO-FOLDING METHODS MOVED BEYOND MEMORISATION?

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
peter.skrinjar@unibas.ch

Jérôme Eberhardt
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
jerome.eberhardt@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
janani.durairaj@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
torsten.schwede@unibas.ch



Practical Cheminformatics Publications Tutorials Blog Videos Resources

Three Papers Demonstrating That Cofolding Still Has a Ways to Go

13 minute read

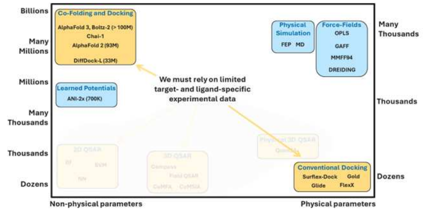
Published: July 21, 2025

Many Posebusters Complexes Have Duplicates Deposited Before 2021

https://patwalters.github.io/Three-Papers-Demonstrating-That-Cofolding-Still-Has-a-Ways-to-Go/?s=03

Near-neighbor effects exist because of the biased manner in which we explore chemical space against biological targets.

Docking: Pure ML vs. Physical Parameters



PoseBusters Benchmark

- Designed to evaluate docking quality on a pharmaceutically relevant set of 308 protein/ligand complexes
- Illustrated quality problems with Pure-ML docking predictions

- M. Buttenschoen, G.M. Morris, C.M. Deane https://doi.org/10.48550/arXiv.2308.05777
- Can be run with a known binding site or as "blind docking"

Known binding-site (pocket-based docking)

- Cognate ligand re-docking
- Top-tier conventional docking methods run by experienced users typically produce 60-80% success at the 2.0 Å RMSD success threshold

Unknown binding-site ("blind" docking)

- Must find the binding sites, dock, and score/rank
- Quite a bit more difficult

Data in **Black** from DiffDock-L paper: https://doi.org/10.48550/arXiv.2402.18396

Method	RMSD $\leq$ 2Å	
Pocket-based docking		
GOLD	58%	
VINA	60%	
DEEPDOCK	20%	
UNI-MOL	22%	
SURFLEX-DOCK	78%	A few dozen parameters

Method	RMSD $\leq$ 2Å	
Blind docking		
EQUIBIND	2%	Millions of parameters
TANKBIND	16%	
DIFFDOCK	38%	Many millions of parameters
ROSETTAFOLD-ALLATOM[†]	42%	
DIFFDOCK-L	50%	
SURFLEX-DOCK	57%	A few dozen parameters

Docking: Pure ML vs. Physical Parameters

DockGen “blind docking” benchmark

DockGen

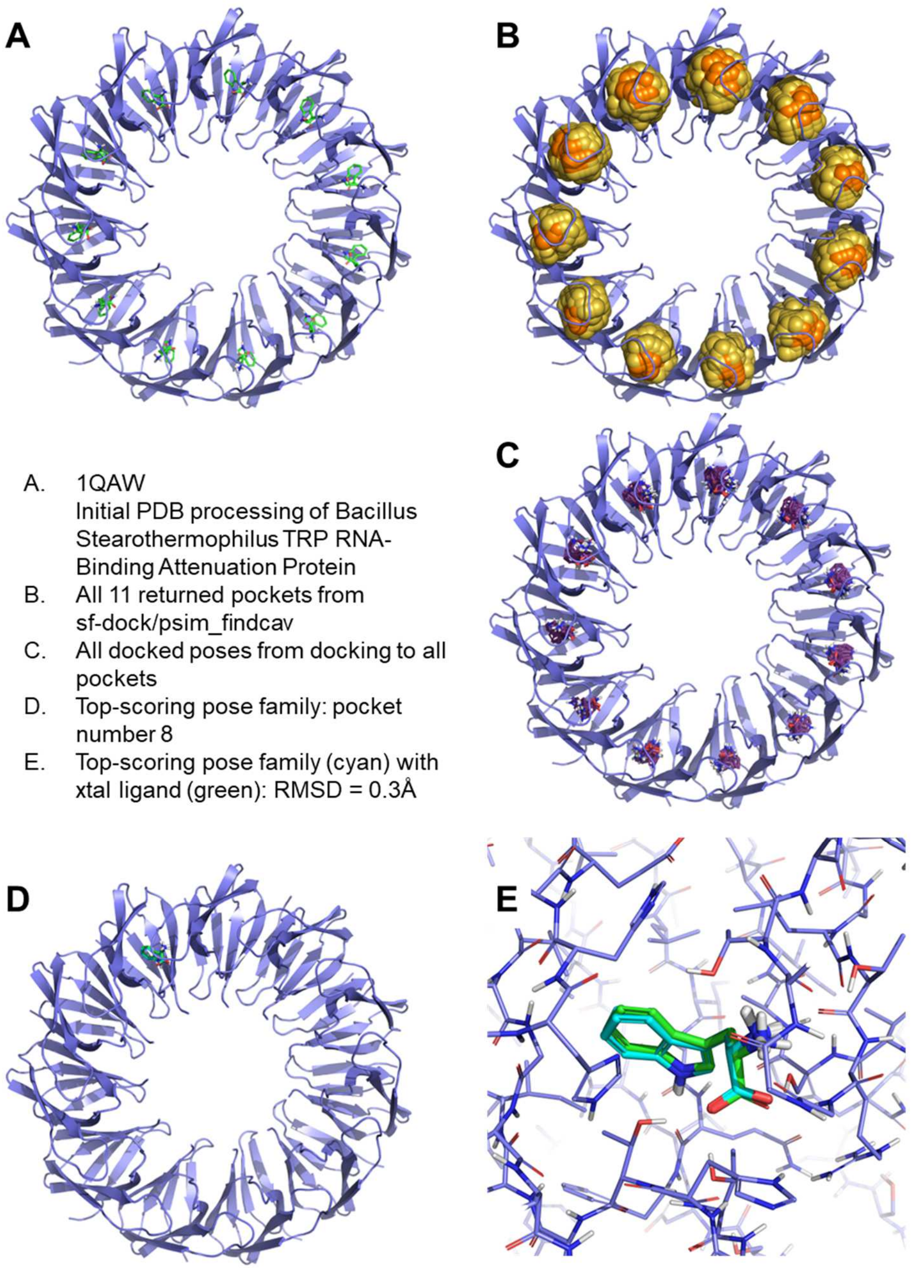
- Designed to contain diverse structures to avoid the problems of near-neighbor effects
- Novel structures compared to PDBBind and BindingMOAD
- Highly diverse set, dominated by ligands that are amino-acids, enzyme co-factors, and metabolites

Pure ML: DiffDock-L

- Number of parameters: **33 million**
- Number of training exemplars: Tens of thousands
- Performance (Å RMSD): **28% < 2.0**, Median = 3.7

Conventional Docking: Surflex-Dock

- Number of parameters: **A few dozen**
- Number of training exemplars: A few hundred (pre-2008)
- Performance (Å RMSD): **41% < 2.0**, Median = 3.3



A. 1QAW
Initial PDB processing of Bacillus Stearothermophilus TRP RNA-Binding Attenuation Protein
B. All 11 returned pockets from sf-dock/psim_findcav
C. All docked poses from docking to all pockets
D. Top-scoring pose family: pocket number 8
E. Top-scoring pose family (cyan) with xtal ligand (green): RMSD = 0.3Å

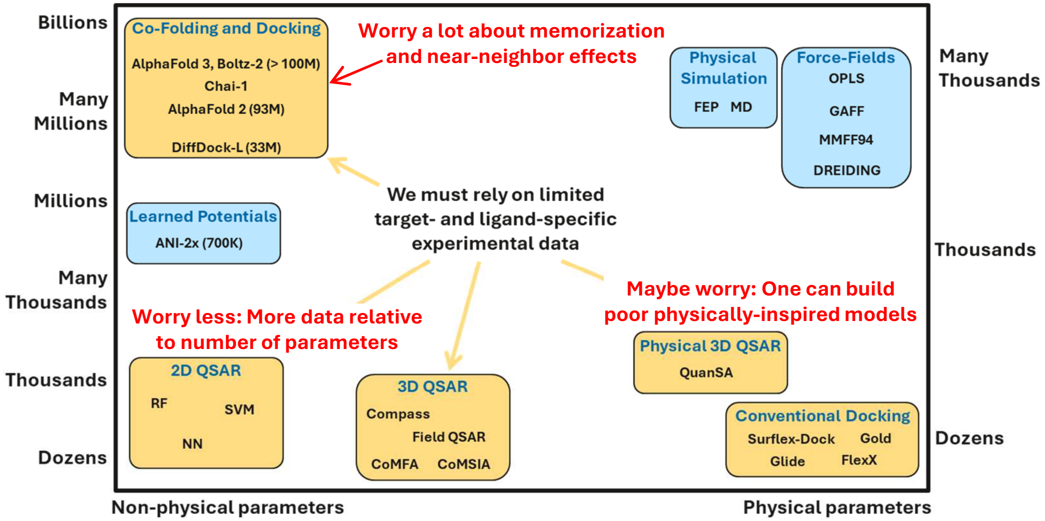
DockGen benchmark from DiffDock-L paper: https://doi.org/10.48550/arXiv.2402.18396

Both Pure ML and Physical Parameter Estimation can succeed

Reliance on limited experimental data to tune millions of parameters is fraught

Pure ML

- Inscrutable black-box parameters that may range into the many millions
- Large models can be highly effective if training data exists that is unbiased and sufficient
- The data in the PDB and ChEMBL required hundreds of thousands of person-years to produce
  - The data are strongly biased
  - Such data will not grow very fast
  - There is no computational method on the horizon that will support accurate data generation



Billions

Many Millions

Millions

Many Thousands

Thousands

Dozens

Co-Folding and Docking

AlphaFold 3, Boltz-2 (> 100M)

Chai-1

AlphaFold 2 (93M)

DiffDock-L (33M)

Worry a lot about memorization and near-neighbor effects

Physical Simulation

FEP MD

Force-Fields

OPLS

GAFF

MMFF94

DREIDING

Many Thousands

Learned Potentials

ANI-2x (700K)

We must rely on limited target- and ligand-specific experimental data

Thousands

Maybe worry: One can build poor physically-inspired models

Worry less: More data relative to number of parameters

Physical 3D QSAR

QuanSA

2D QSAR

RF SVM NN

3D QSAR

Compass

Field QSAR

CoMFA CoMSIA

Conventional Docking

Surflex-Dock Gold

Glide FlexX

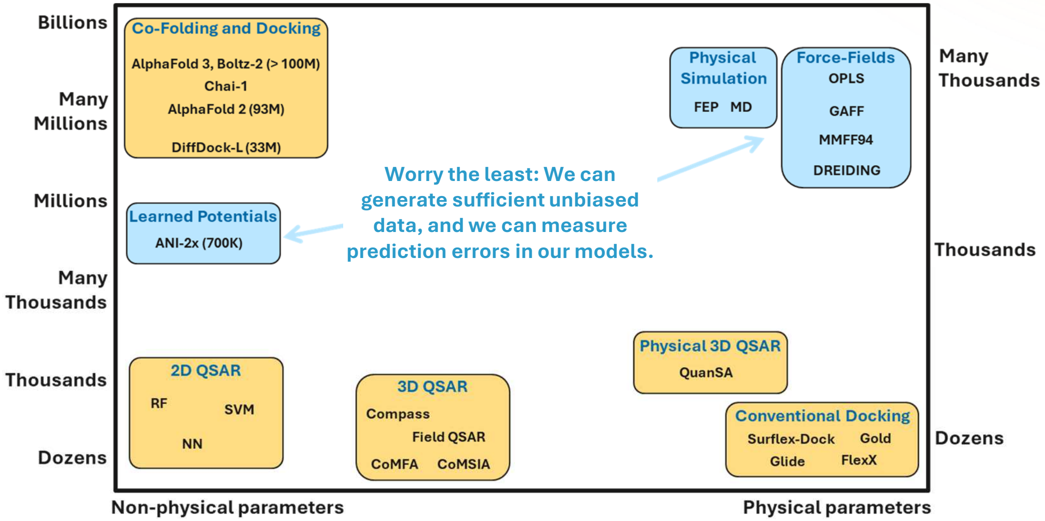
Dozens

Non-physical parameters

Physical parameters

Physical Parameter Estimation

- Models that have parameters which mirror a physically sensible understanding of underlying physics have a built-in advantage for generalization
  - They lean toward being *causally-based*, which ameliorates dependency on the central ML assumption
- There is still wide variation in the quality of such models
  - However, the best-performing of such approaches often exhibit substantially better predictive behavior than large Pure-ML models that rely on limited/biased experimental data



Billions

Many Millions

Millions

Many Thousands

Thousands

Dozens

Co-Folding and Docking

AlphaFold 3, Boltz-2 (> 100M)

Chai-1

AlphaFold 2 (93M)

DiffDock-L (33M)

Physical Simulation

FEP MD

Force-Fields

OPLS

GAFF

MMFF94

DREIDING

Many Thousands

Learned Potentials

ANI-2x (700K)

Worry the least: We can generate sufficient unbiased data, and we can measure prediction errors in our models.

Thousands

Physical 3D QSAR

QuanSA

2D QSAR

RF SVM NN

3D QSAR

Compass

Field QSAR

CoMFA CoMSIA

Conventional Docking

Surflex-Dock Gold

Glide FlexX

Dozens

Non-physical parameters

Physical parameters

Acknowledgements

Key collaborators

- Optibrium
  - Himani Tandon
  - Andrew Smith
  - Marietta Homor
  - Irena Kiso
  - Matt Segall
- BMS
  - Alex Brueckner
  - Luciano Mueller
  - Christine Jorge
  - Purnima Khandelwal
  - Janet Caceres-Cortes
  - Stephen Johnson
- Merck
  - Ed Sherer
  - Mikhail Reibarkh
  - Qi Gao
  - Charles Lesburg
- Relay
  - Pat Walters
  - Dimitri Moustakas
- Carnegie Mellon
  - Olexandr Isayev

BioPharmics Platform v5.1: Linux, Windows, Mac

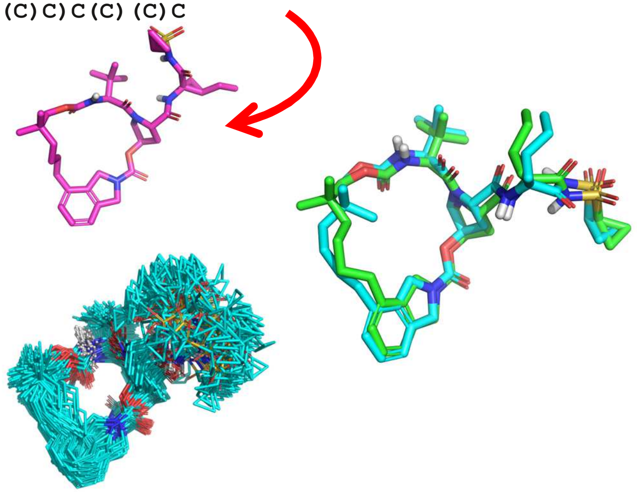
- Tools
  - ForceGen 2D to 3D and conformer generation
  - Fast, accurate macrocycle elaboration
- Docking
  - Class-leading docking solution for pose prediction and virtual screening
  - Large-scale PDB processing
  - Protein binding site comparison, alignment, and selection
- X-Ray
  - xGen real-space fitting of ligands into X-ray density
  - De novo ligand fitting
  - Macrocycles and non-macrocycles
- Similarity
  - eSim: Electrostatic field and surface-based similarity method
  - Virtual screening and scaffold replacement
  - Multiple ligand alignment
- Affinity
  - QuanSA: Unique solution to the 3D QSAR problem
  - Rigorous solution to the alignment problem using multiple-instance machine learning
  - Scaffold independent extrapolative prediction
  - Rapid application to candidate molecules

BioPharmics™

a division of optibrium

ForceGen

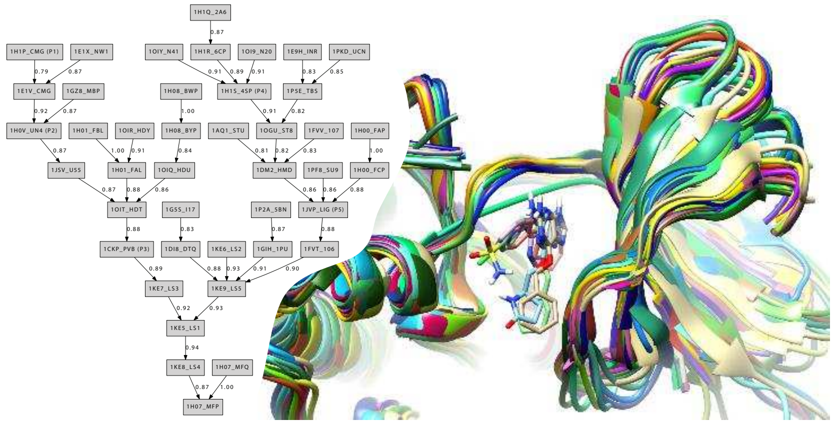
```
CC[C@@H]1C[C@@]1(C(=O)NS(=O)(=O)C2CC2)
NC(=O)[C@@H]3C[C@@H]4CN3C(=O)[C@@H](NC
(=O)OCC(CCCCC5=C6CN(CC6=CC=C5)C(=O)O4)
(C)C)C(C)(C)C
```

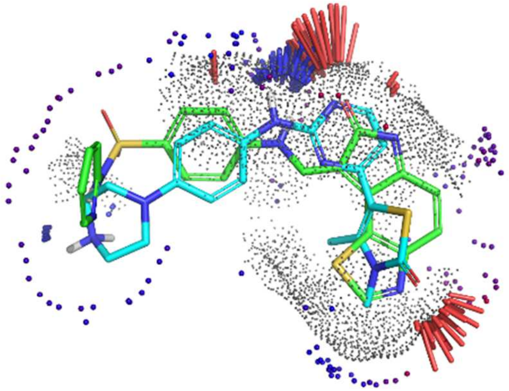


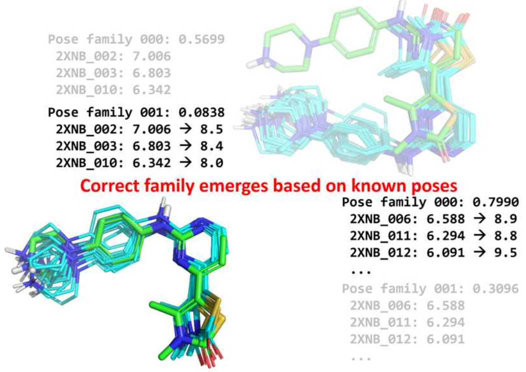
BioPharmics™

A platform for 3D molecular design

Surflex-Dock





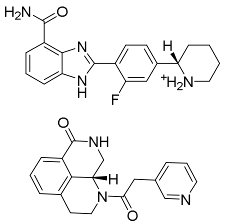


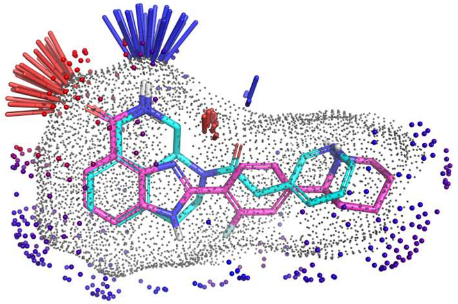
```
Pose family 000: 0.5699
 2XNB_002: 7.006
 2XNB_003: 6.803
 2XNB_010: 6.342

Pose family 001: 0.0838
 2XNB_002: 7.006 → 8.5
 2XNB_003: 6.803 → 8.4
 2XNB_010: 6.342 → 8.0
```

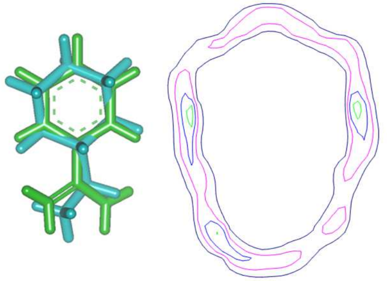
Correct family emerges based on known poses

```
Pose family 000: 0.7990
 2XNB_006: 6.588 → 8.9
 2XNB_011: 6.294 → 8.8
 2XNB_012: 6.091 → 9.5
 ...
Pose family 001: 0.3096
 2XNB_006: 6.588
 2XNB_011: 6.294
 2XNB_012: 6.091
 ...
```

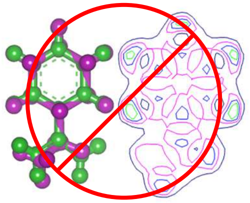




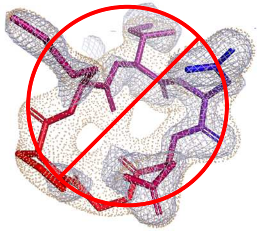
Surface-distance differences

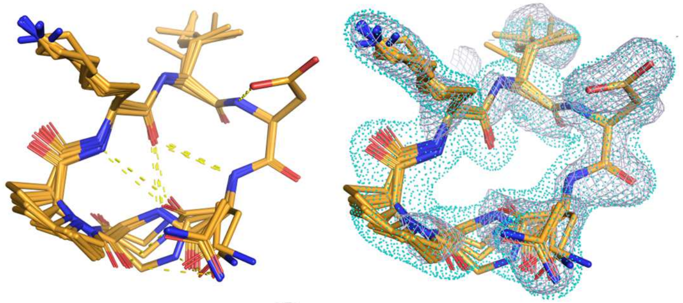


Atom-centered Gaussians

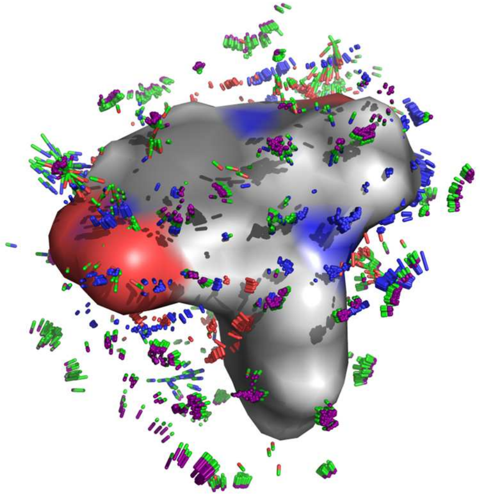


eSim

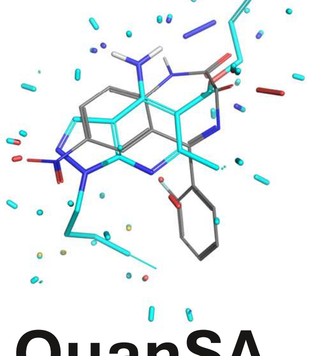




xGen



Pred pK_i = **8.0** (8.3)



QuanSA