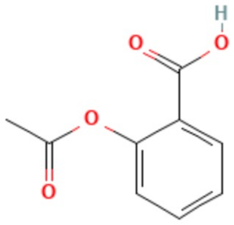
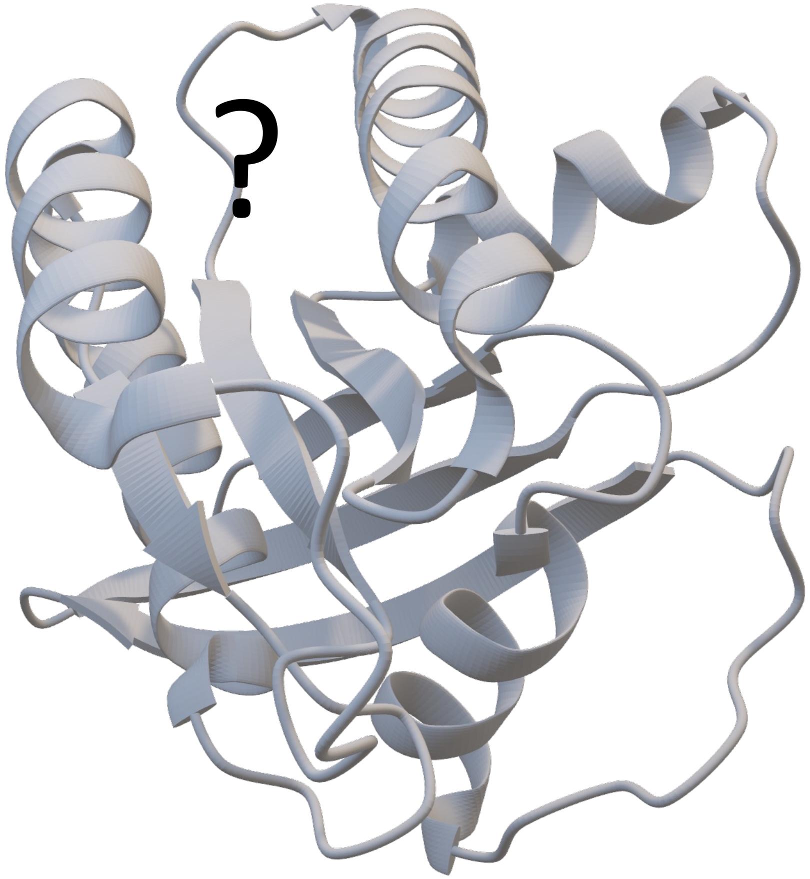
Drug Discovery using Machine Learning



CC(=O)OC1=CC=CC=C1C(=O)O

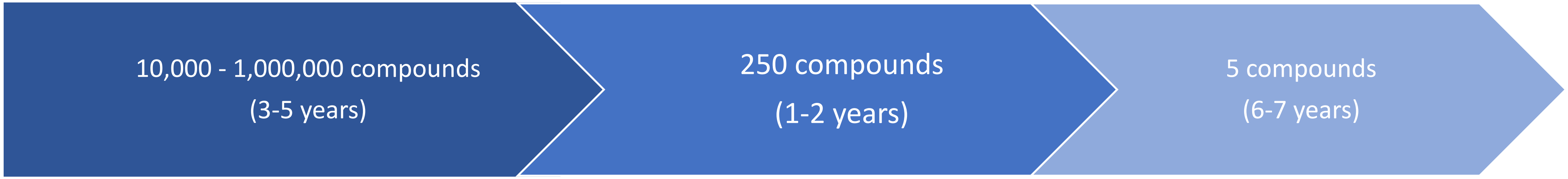


Drug discovery pipeline in a nutshell

Hit identification

Lead optimization

Clinical phase



10,000 - 1,000,000 compounds
(3-5 years)

250 compounds
(1-2 years)

5 compounds
(6-7 years)

Using experimental high-throughput screening to identify lead compounds:

- Low success rate (drug space ca. 10^{60})
- High entry barrier:
    1. Large compound library
    2. Expensive experiments

Optimization of chemical properties:

- Toxicity
- Tumor growth inhibition
- Selectivity against other proteins

Machine learning has the potential to optimize and improve the screening stage to **reduce time & cost** and **increase success rate** in the clinical phase.

Credits: https://www.researchgate.net/publication/308045230_Omics-Informed_Drug_and_Biomarker_Discovery_Opportunities_Challenges_and_Future_Perspectives/figures?lo=1

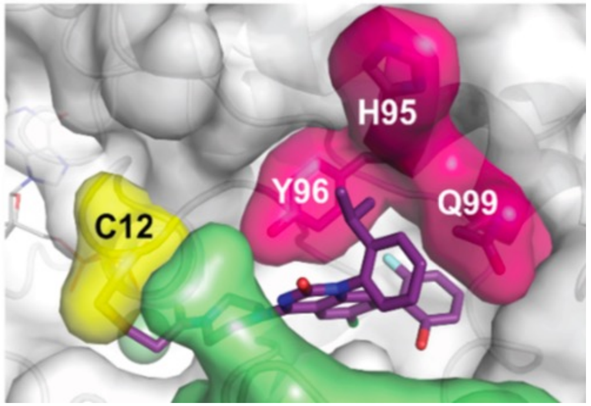
Case study: KRAS G12C

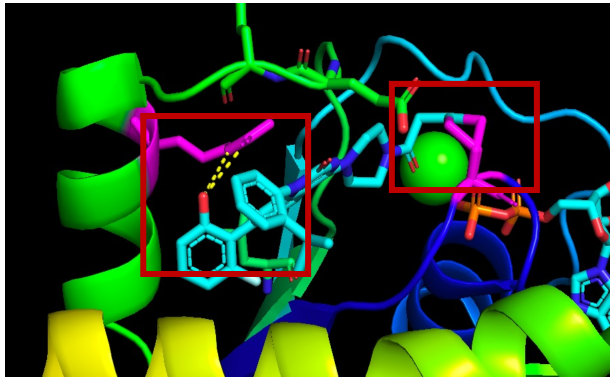
Goal: Design a compound that is effectively interacting with the KRAS protein.

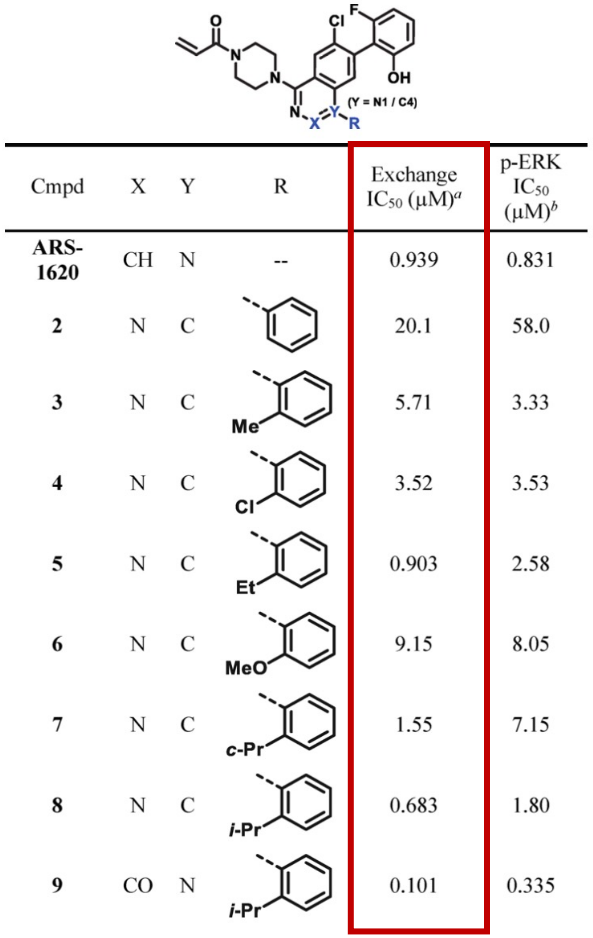
Development pipeline from Amgen:

1. Screen experimentally a 300k compound library
2. Two hits (potent) compounds were identified
3. Optimize lead compound by decreasing IC50 (affinity)

The 3D structure contains information about residue interactions contributing to the affinity.






O
N
N
Cl
F
OH
(Y = N1 / C4)
N
X
Y
R

Cmpd	X	Y	R	Exchange IC$_{50}$ (μM)^a	p-ERK IC$_{50}$ (μM)^b
ARS-1620	CH	N	--	0.939	0.831
2	N	C		20.1	58.0
3	N	C	Me	5.71	3.33
4	N	C	Cl	3.52	3.53
5	N	C	Et	0.903	2.58
6	N	C	MeO	9.15	8.05
7	N	C	*c*-Pr	1.55	7.15
8	N	C	*i*-Pr	0.683	1.80
9	CO	N	*i*-Pr	0.101	0.335

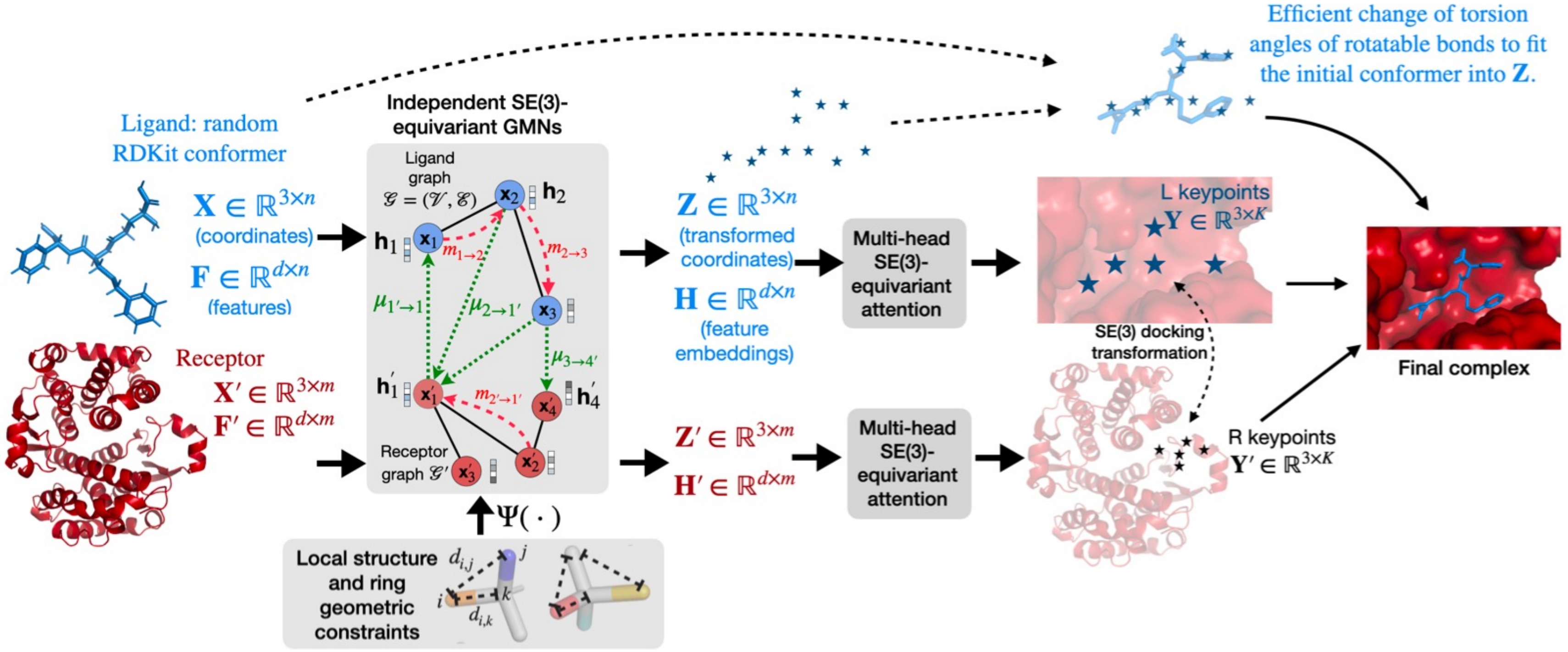
Credits: Lanman et al.
https://doi.org/10.1021/acs.jmedchem.9b01180

EquiBind – Predicting the 3D protein-ligand complex

Overview:

1. Compute cheap initial 3D conformation of the compound
2. Embed protein & compound
3. Predict rotation & transformation of the compound

Assumption: Fixed protein structure


Ligand: random RDKit conformer
$\mathbf{X} \in \mathbb{R}^{3 \times n}$ (coordinates)
$\mathbf{F} \in \mathbb{R}^{d \times n}$ (features)
Receptor
$\mathbf{X}' \in \mathbb{R}^{3 \times m}$
$\mathbf{F}' \in \mathbb{R}^{d \times m}$
Independent SE(3)-equivariant GMNs
Ligand graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
$\mathbf{h}_1$
$\mathbf{h}_2$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$m_{1\to2}$
$m_{2\to3}$
$\mu_{1'\to1}$
$\mu_{2\to1'}$
$\mu_{3\to4'}$
$\mathbf{h}'_1$
$\mathbf{h}'_4$
$\mathbf{x}'_1$
$\mathbf{x}'_2$
$\mathbf{x}'_3$
$\mathbf{x}'_4$
$m_{2'\to1'}$
Receptor graph $\mathcal{G}'$
$\Psi(\cdot)$
Local structure and ring geometric constraints
$d_{i,j}$
$d_{i,k}$
i
j
k
$\mathbf{Z} \in \mathbb{R}^{3 \times n}$ (transformed coordinates)
$\mathbf{H} \in \mathbb{R}^{d \times n}$ (feature embeddings)
$\mathbf{Z}' \in \mathbb{R}^{3 \times m}$
$\mathbf{H}' \in \mathbb{R}^{d \times m}$
Multi-head SE(3)-equivariant attention
Multi-head SE(3)-equivariant attention
L keypoints
$\mathbf{Y} \in \mathbb{R}^{3 \times K}$
SE(3) docking transformation
R keypoints
$\mathbf{Y}' \in \mathbb{R}^{3 \times K}$
Efficient change of torsion angles of rotatable bonds to fit the initial conformer into $\mathbf{Z}$.
Final complex

Credits image: Stärk et al. - http://arxiv.org/abs/2202.05146

Equivariant Graph Neural Network

Input:

- Compound graph $(\mathcal{V}, \mathcal{E})$ with $X \in \mathbb{R}^{3 \times n}$ coordinates and $\mathrm{H} \in \mathbb{R}^{d \times n}$ features
- Receptor graph $(\mathcal{V}', \mathcal{E}')$ with $X' \in \mathbb{R}^{3 \times m}$ coordinates and $H' \in \mathbb{R}^{d \times m}$ features

Notation:

- $\varphi^{e,h}$ feed-forward NN with d-dimensional output & φ^{x} feed-forward NN with scalar output
- $a_{\cdot \to \cdot}$ attention coefficient
- $f_{\cdot \to \cdot}$ edge features, e.g. bond type

Single Layer (IEGMN):

1. Compute edge feature

$$\mathbf{m}_{j\to i} = \varphi^e(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \|\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}\|^2, \mathbf{f}_{j\to i}), \forall (i,j) \in \mathcal{E} \cup \mathcal{E}'$$

$$\mu_{j'\to i} = a_{j'\to i}\mathbf{W}\mathbf{h}_{j'}^{(l)}, \forall i \in \mathcal{V}, j' \in \mathcal{V}' \text{ or } i \in \mathcal{V}', j' \in \mathcal{V}$$

2. Aggregate over nodes

$$\mathbf{m}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{j\to i}, \forall i \in \mathcal{V} \cup \mathcal{V}'$$

$$\mu_i = \sum_{j' \in \mathcal{V}'} \mu_{j'\to i}, \forall i \in \mathcal{V}, \quad \text{and} \quad \mu_i' = \sum_{j \in \mathcal{V}} \mu_{j \to i'}, \forall i \in \mathcal{V}'$$

3. Update node features

$$\mathbf{x}_i^{(l+1)} = \Psi\left(\mathbf{x}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \frac{\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}}{\|\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}\|} \varphi^x(\mathbf{m}_{j\to i})\right)$$

$$\mathbf{h}_i^{(l+1)} = (1-\beta)\cdot \mathbf{h}_i^{(l)} + \beta \cdot \varphi^h(\mathbf{h}_i^{(l)}, \mathbf{m}_i, \mu_i, \mathbf{f}_i), \forall i \in \mathcal{V} \cup \mathcal{V}'$$

Credits: Stärk et al. - http://arxiv.org/abs/2202.05146

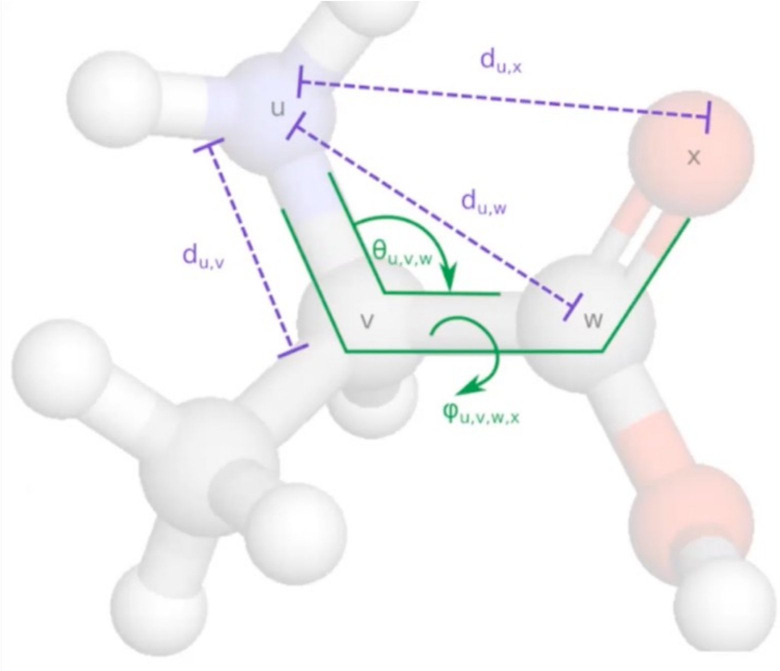
Enforcing a chemical plausible geometry

Motivation: Local atom structures (e.g. bond length & adjacent bond angels) are mostly rigid.

⬇

Minimize the loss S for a chemical plausible conformer C to enforce LAS

$$\begin{aligned}\mathcal{S}(X,C) = &\sum_{(i,j)\in\mathcal{E}} (d_C^2(i,j) - d_X^2(i,j))^2 \\ &+ \sum_{(i,j):\ \text{2-hops away in } \mathcal{G}} (d_C^2(i,j) - d_X^2(i,j))^2 \\ &+ \sum_{(i,j):\ i \text{ in aromatic ring with } j} (d_C^2(i,j) - d_X^2(i,j))^2\end{aligned}$$



with gradient descent

$$\Psi(X) = \Psi_T \circ \cdots \circ \Psi_1(X), \quad \Psi_t(X) = X - \eta\nabla_X \mathcal{S}(X,C), \forall t$$

Credits: Stärk et al. - http://arxiv.org/abs/2202.05146

Kabsch algorithm: Finding the right rotation and translation

1. Identify K keypoints for receptor and compound: $Y' \in \mathbb{R}^{K\times 3}$ & $Y \in \mathbb{R}^{K\times 3}$ with $y_k = \sum_{i=1}^{n} \alpha_i^k x_i^L$

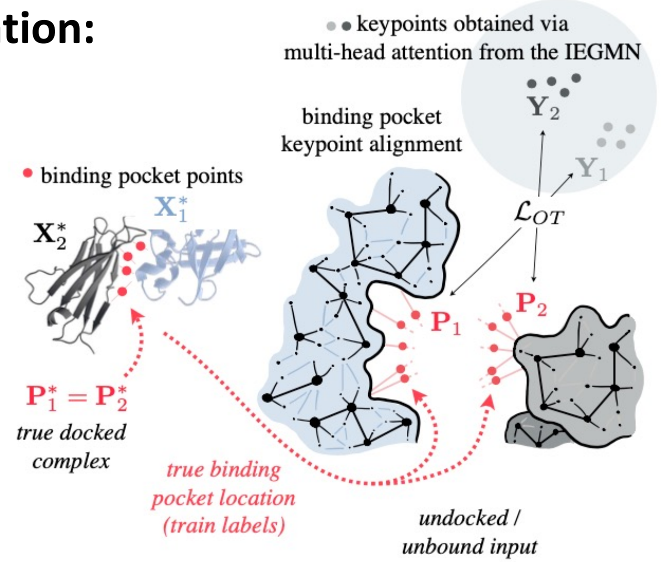
2. Compute rotation and translation (Kabsch algorithm)

3. MSE loss: $\tilde{X} = RX + t$ ➡ $\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^* - \tilde{\mathbf{x}}_i\right\|^2$

4. Enforce the keypoints to be binding pocket points of the compound and receptor with **optimal transport loss**

$$\mathcal{L}_{\text{OT}} = \min_{\mathbf{T}\in\mathcal{U}(S,K)} \langle \mathbf{T}, \mathbf{C} \rangle, \quad \text{where } \mathbf{C}_{s,k} = \left\|\mathbf{y}_{1k} - \mathbf{p}_{1s}\right\|^2 + \left\|\mathbf{y}_{2k} - \mathbf{p}_{2s}\right\|^2,$$

Motivation:



Kabsch algorithm:

1. $A = Y'Y^T \in \mathbb{R}^{3\times 3}$
2. SVD: $A = U_2 S U_1^T$
3. Rotation: $\mathrm{R} = U_2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} U_1^T$, where $d = \text{sign}(\det(U_2 U_1^T))$
4. Translation: $t = \mu(Y') - R\mu(Y)$

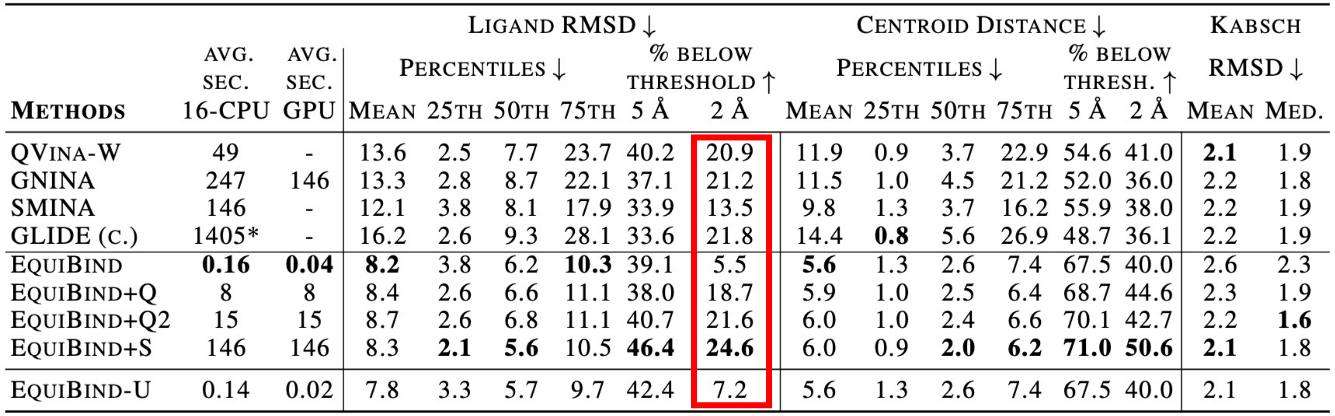
Reference: S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns“, 1991

Credits: Ganea et al. - http://arxiv.org/abs/2111.07786

Results & What is missing?

Results:

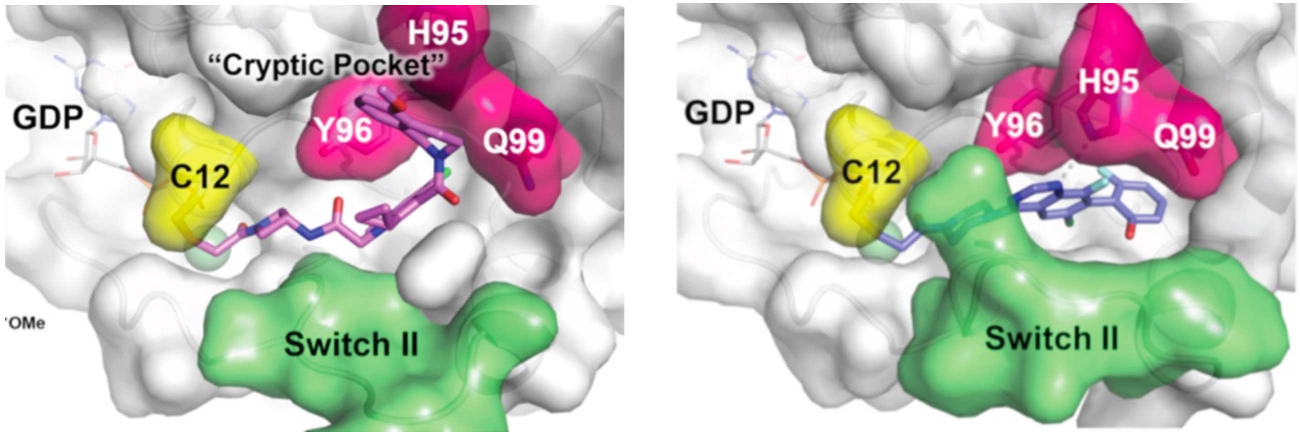
- EquiBind is significantly faster
- Combining it with finetuning method helps a lot (EquiBind + S)
- Low percentage of RMSD below 2A (bad ☹)



			Ligand RMSD ↓						Centroid Distance ↓						Kabsch	
	Avg.	Avg.	Percentiles ↓				% below		Percentiles ↓				% below		RMSD ↓	
	sec.	sec.					threshold ↑						thresh. ↑			
Methods	16-CPU	GPU	Mean	25th	50th	75th	5 Å	2 Å	Mean	25th	50th	75th	5 Å	2 Å	Mean	Med.
QVina-W	49	-	13.6	2.5	7.7	23.7	40.2	20.9	11.9	0.9	3.7	22.9	54.6	41.0	**2.1**	1.9
GNINA	247	146	13.3	2.8	8.7	22.1	37.1	21.2	11.5	1.0	4.5	21.2	52.0	36.0	2.2	1.8
SMINA	146	-	12.1	3.8	8.1	17.9	33.9	13.5	9.8	1.3	3.7	16.2	55.9	38.0	2.2	1.9
GLIDE (c.)	1405*	-	16.2	2.6	9.3	28.1	33.6	21.8	14.4	**0.8**	5.6	26.9	48.7	36.1	2.2	1.9
EquiBind	**0.16**	**0.04**	**8.2**	3.8	6.2	**10.3**	39.1	5.5	**5.6**	1.3	2.6	7.4	67.5	40.0	2.6	2.3
EquiBind+Q	8	8	8.4	2.6	6.6	11.1	38.0	18.7	5.9	1.0	2.5	6.4	68.7	44.6	2.3	1.9
EquiBind+Q2	15	15	8.7	2.6	6.8	11.1	40.7	21.6	6.0	1.0	2.4	6.6	70.1	42.7	2.2	**1.6**
EquiBind+S	146	146	8.3	**2.1**	**5.6**	10.5	**46.4**	**24.6**	6.0	0.9	**2.0**	**6.2**	**71.0**	**50.6**	**2.1**	1.8
EquiBind-U	0.14	0.02	7.8	3.3	5.7	9.7	42.4	7.2	5.6	1.3	2.6	7.4	67.5	40.0	2.1	1.8

What is missing?

- Binding affinity (IC50)
- Protein is not always rigid



Change of protein pocket due to different compound

Backup

Can we trust the initial conformer?

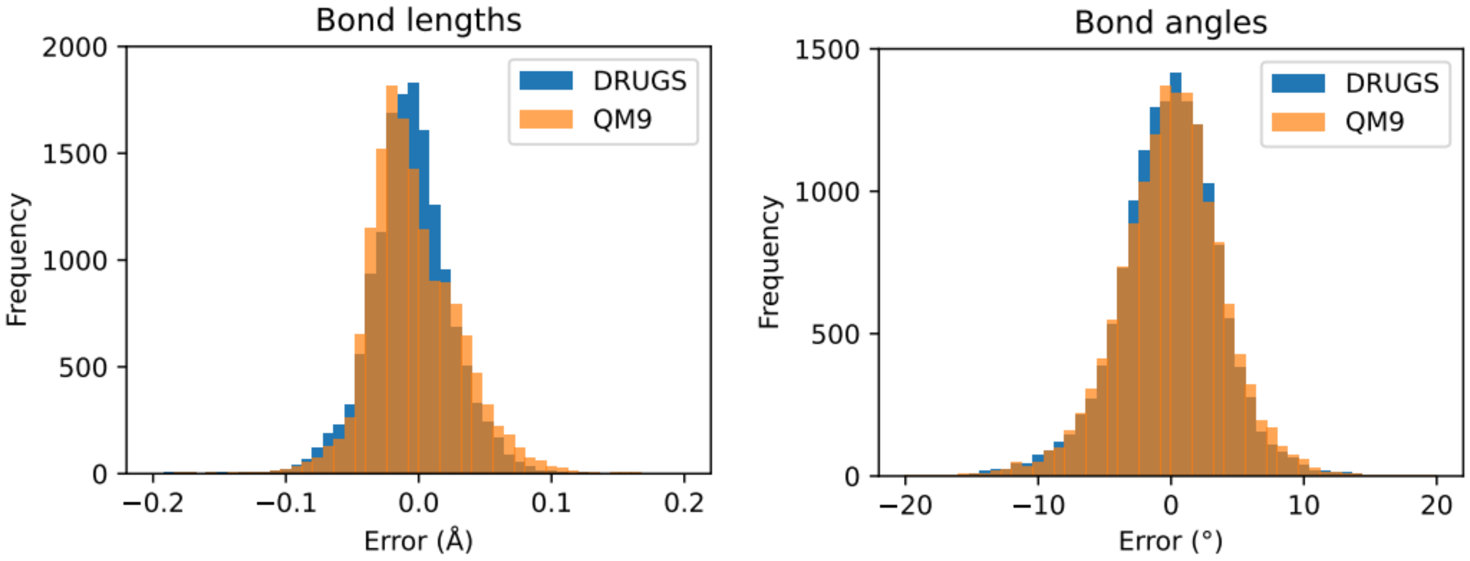


Figure 5: Histogram of the errors in 15000 predicted bond lengths and angles from randomly sampled molecules in GEOM-DRUGS and GEOM-QM9.

Credits: https://arxiv.org/pdf/2206.01729.pdf