

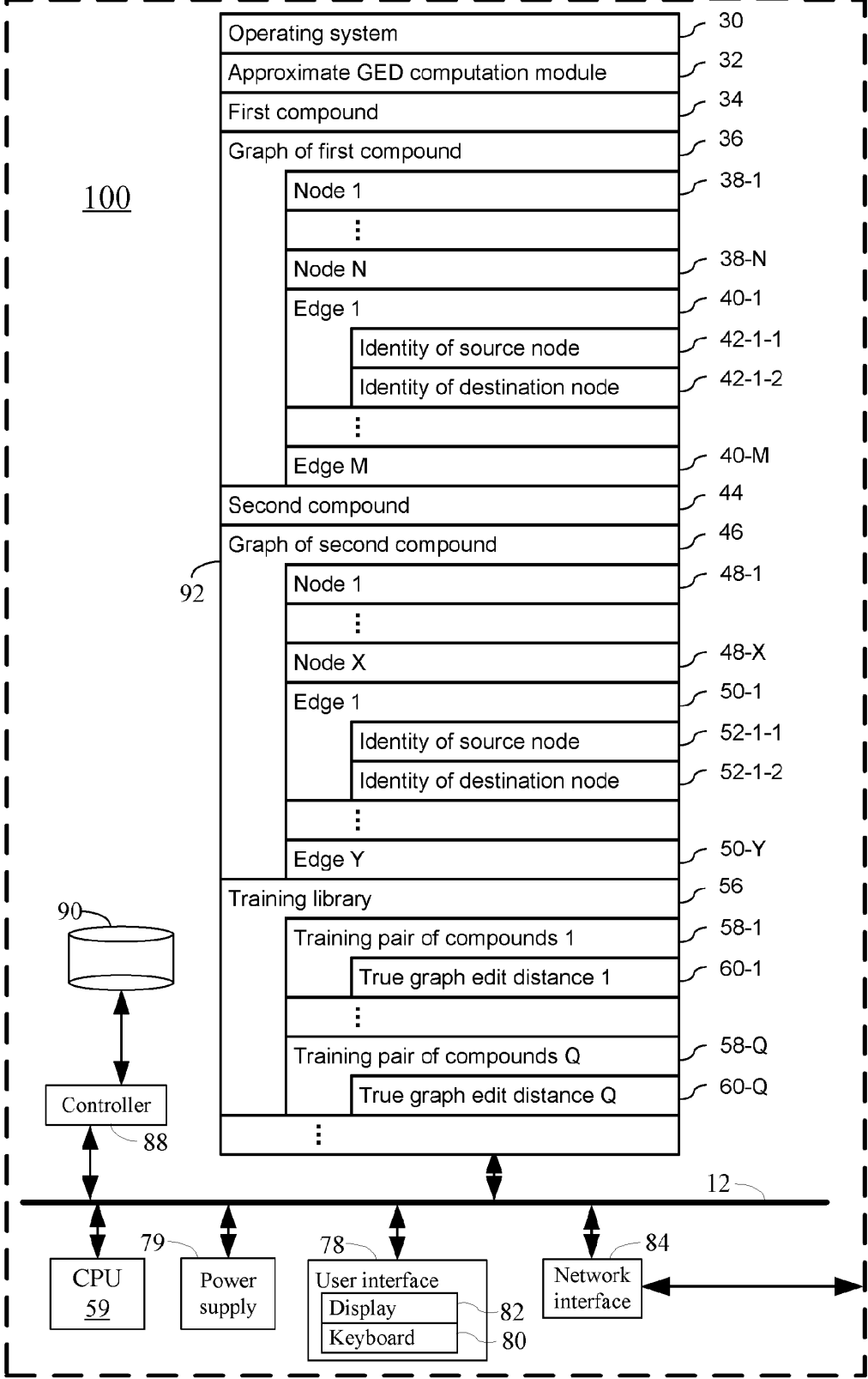


Fig. 1

(200) Systems and methods for estimating graph edit distance between a first compound and a second compound are provided.

(202) A first graph is obtained that represents the first compound. The first graph comprises a first plurality of nodes and a first plurality of edges. Each atom of the first compound is represented by a node in the first plurality of nodes. Each bond of the first compound is represented by an edge in the first plurality of edges.

(204) A second graph is obtained representing the second compound. The second graph comprises a second plurality of nodes and a second plurality of edges. Each atom of the second compound is represented by a node in the second plurality of nodes. Each bond of the second compound is represented by an edge in the second plurality of edges.

(206) The first compound and the second compound each satisfy two or more rules, three or more rules, or all four rules of the Lipinski's rule of Five: (i) not more than five hydrogen bond donors, (ii) not more than ten hydrogen bond acceptors, (iii) a molecular weight under 500 Daltons, and (iv) a LogP under 5.

(208) The first compound and the second compound have an exact graph edit distance of between 1 and 10.

(210) The first compound and the second compound have an exact graph edit distance of between 1 and 15.

(212) The first compound and the second compound are each an organic compound having a molecular weight of less than 500 Daltons, less than 1000 Daltons, less than 2000 Daltons, less than 4000 Daltons, less than 6000 Daltons, less than 8000 Daltons, less than 10000 Daltons, or less than 20000 Daltons.

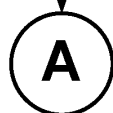


Fig. 2A

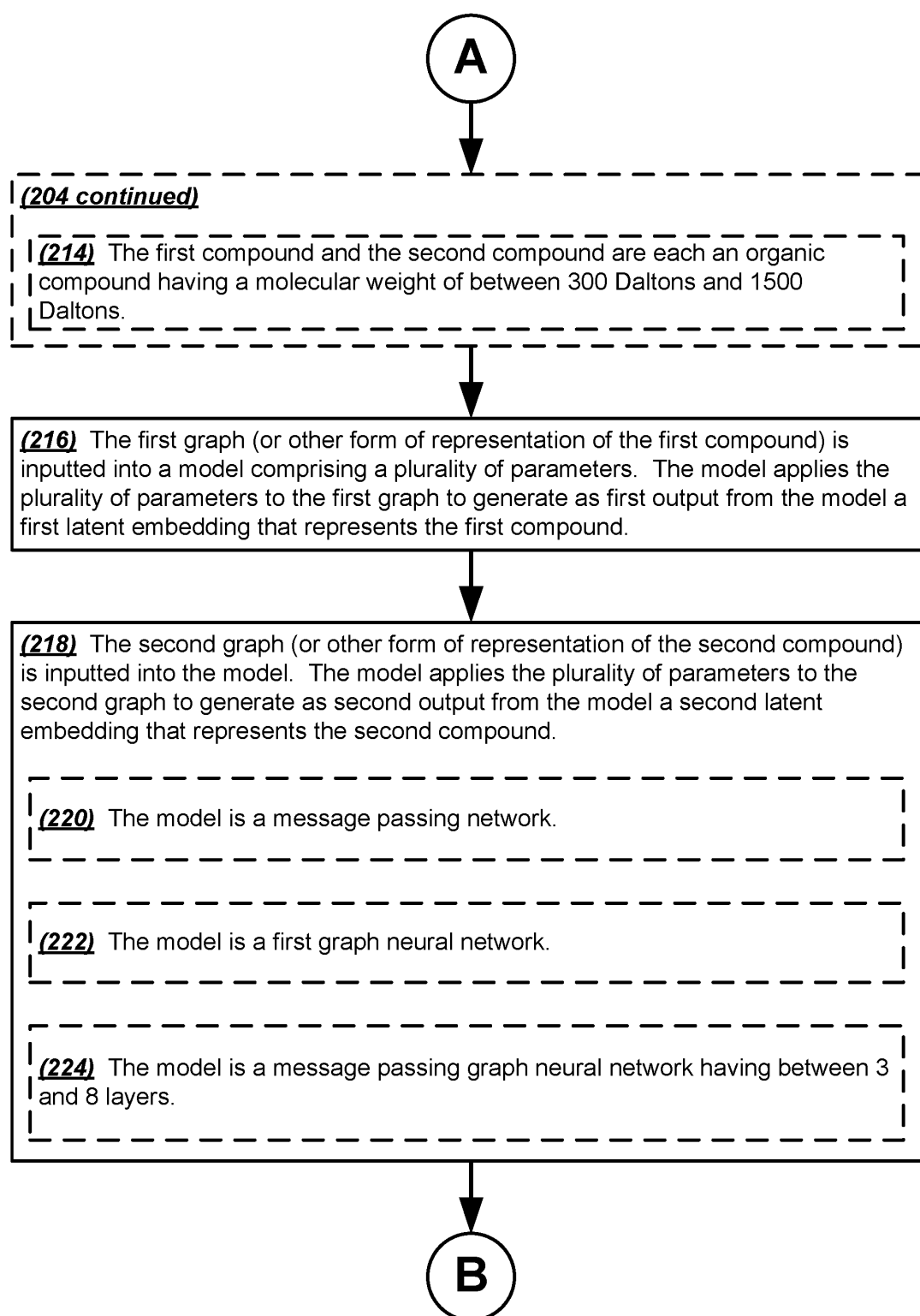


Fig. 2B

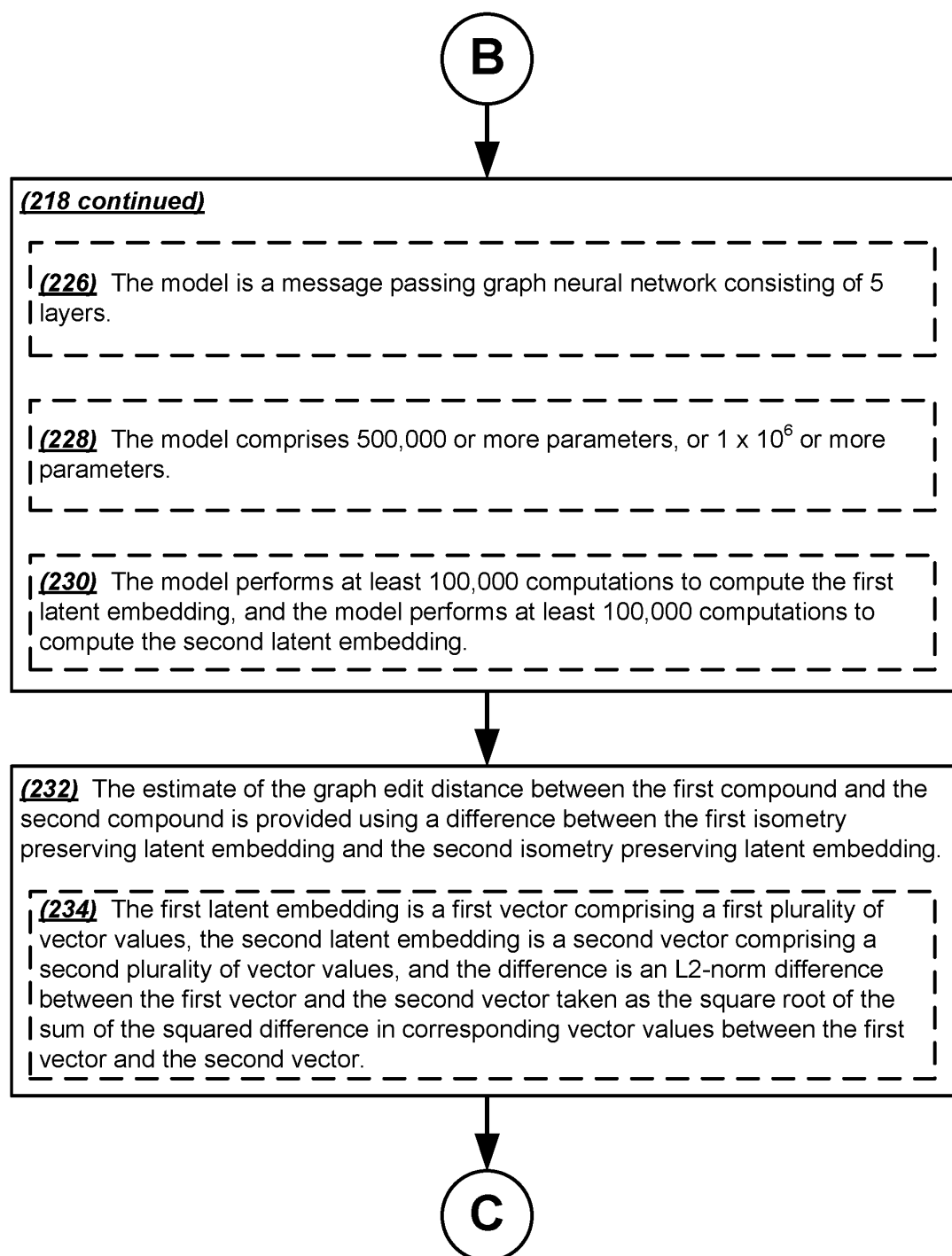
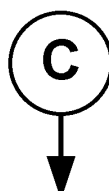


Fig. 2C

**(232 continued)**

(236) A determination is made as to whether the second compound is a chemically relevant analog of the first compound when the estimate of the graph edit distance falls below a threshold value.

(238) The obtaining the first graph, obtaining the second graph, inputting the first graph, inputting the second graph, and providing the approximate GED is repeated for each respective unique pair of compounds in a database of compounds, where a first compound in the respective unique pair of compounds is the first compound and a second compound in the respective unique pair of compounds is the second compound, thereby forming a plurality of graph edit distances. The plurality of graph edit distances are then clustered into a plurality of clusters by evaluating the plurality of distances with a criterion function. Then, compounds are selected from one or more clusters in the plurality of clusters for evaluation in a drug discovery program.

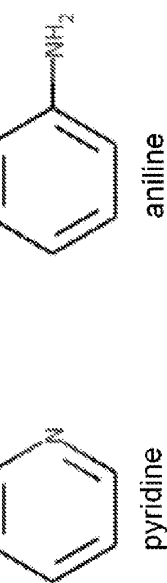
(240) The estimate of the graph edit distance is used to build a quantitative structure activity relationship model that includes the first compound and the second compound.

(242) The estimate of the graph edit distance is used to determine an activity cliff with respect to a drug target.

Fig. 2D

What is graph edit distance (GED)?

GED is the minimum edit distance among all paths connecting two graphs



- * Molecules can be represented as graphs
 - * Atoms -> nodes
 - * Bonds -> edges
- * Graph edits consists of 6 elementary operations
 1. N-> C substitution
 2. Addition of N
 3. Addition of C-N bond
- * Each elementary operation is associated with a cost
- * Path is defined as a sequence of elementary operations to transform 1 graph to another
- * Edit distance for a given path is the sum of costs of operations in the path
- * GED is the minimum edit distance among all possible paths

Fig. 3A
(Prior Art)

Training and testing dataset

Non-trivial train/test pair split created using Taylor-Butina clustering

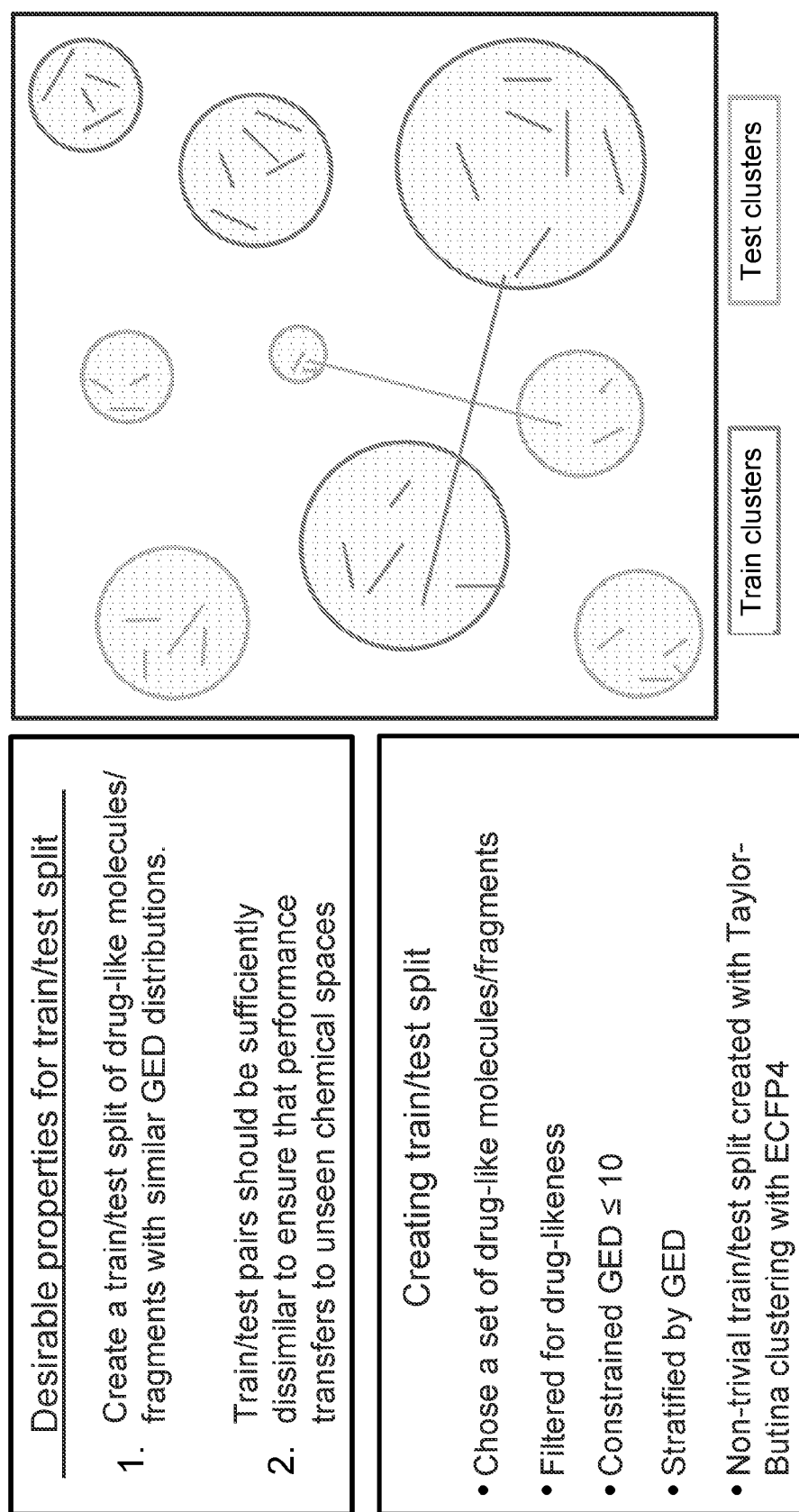


Fig. 3B

Training and test dataset
Drug-like dataset is more challenging than the fragment dataset

	Fragment dataset	Drug-like dataset
	Drug-like molecules cut up into fragments	Drug-like molecules from ChEMBL 30
GED type	Exact GED	Chemical GED
How GED was computed?	Depth-first GED	Commercial software
Avg # heavy atoms	13.7	19.6
# unique	2.9K	47K
Train/test set size	50K/5K	45K/4.5K

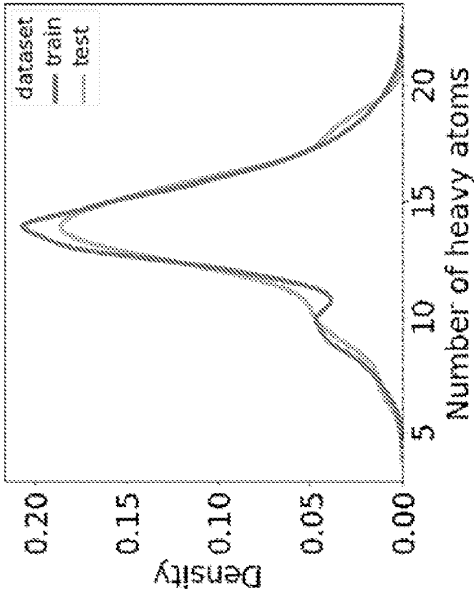


Fig. 4

Performance of SimGNN

	Fragment dataset	Drug-like dataset
MSE	2.93	2.91
Spearman's ρ	0.84	0.80
Kendall's τ_B	0.66	0.60

- Predictions are lacking accuracy with high MSE
- Ranking performance deteriorates on drug-like dataset

Fig. 5

Formal properties of GED

SimGNN prediction violates properties of a distance metric

GED is a distance metric and should satisfy the properties of distance metrics

1. Distance to self is 0: $d(x, x) = 0$
2. Symmetry: $d(x, y) = d(y, x)$
3. Positivity: $d(x, y) > 0$ if $x \neq y$
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

SimGNN predictions violate properties of a distance metric

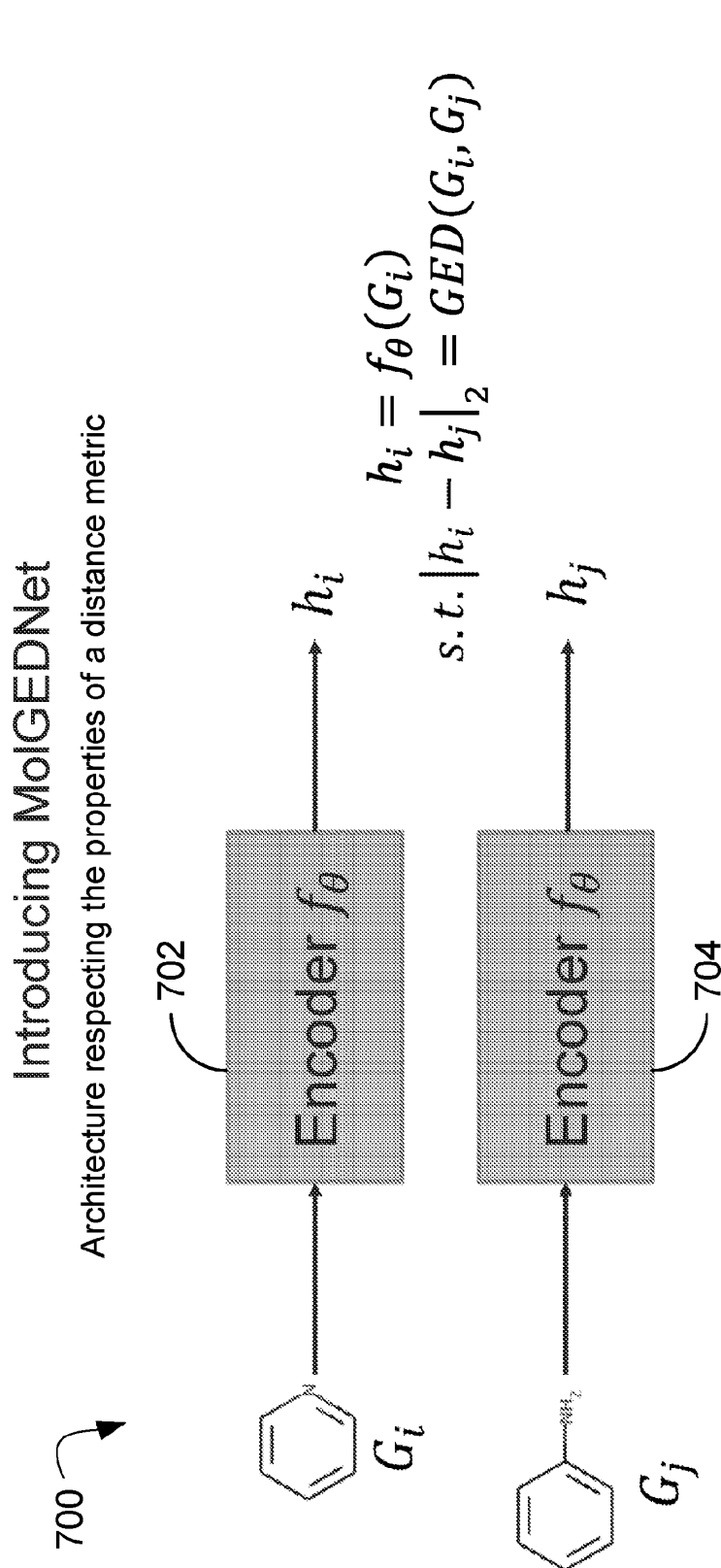
Symmetry violations:

$$|\hat{d}(x, y) - \hat{d}(y, x)|^2 = 0.67$$

Distance to self

$$|\hat{d}(x, x)|^2 = 1.52$$

Fig. 6

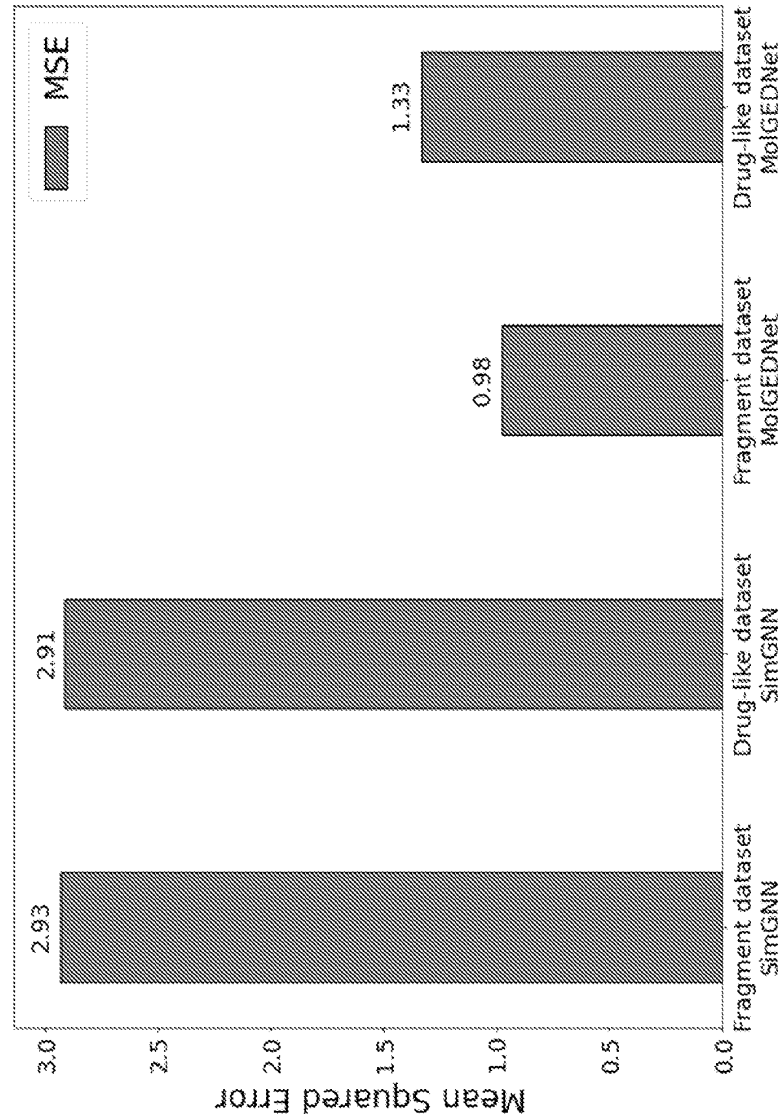


- Encoder f_θ (e.g., parameters of a MPNN) trained such that the distance in latent space is equal to the GED between molecules
- Properties of a distance metric are satisfied by construction
 - Positivity axiom is limited by the ability of model 702 / 704 to distinguish the two graphs

Fig. 7

Performance of MolGEDNet

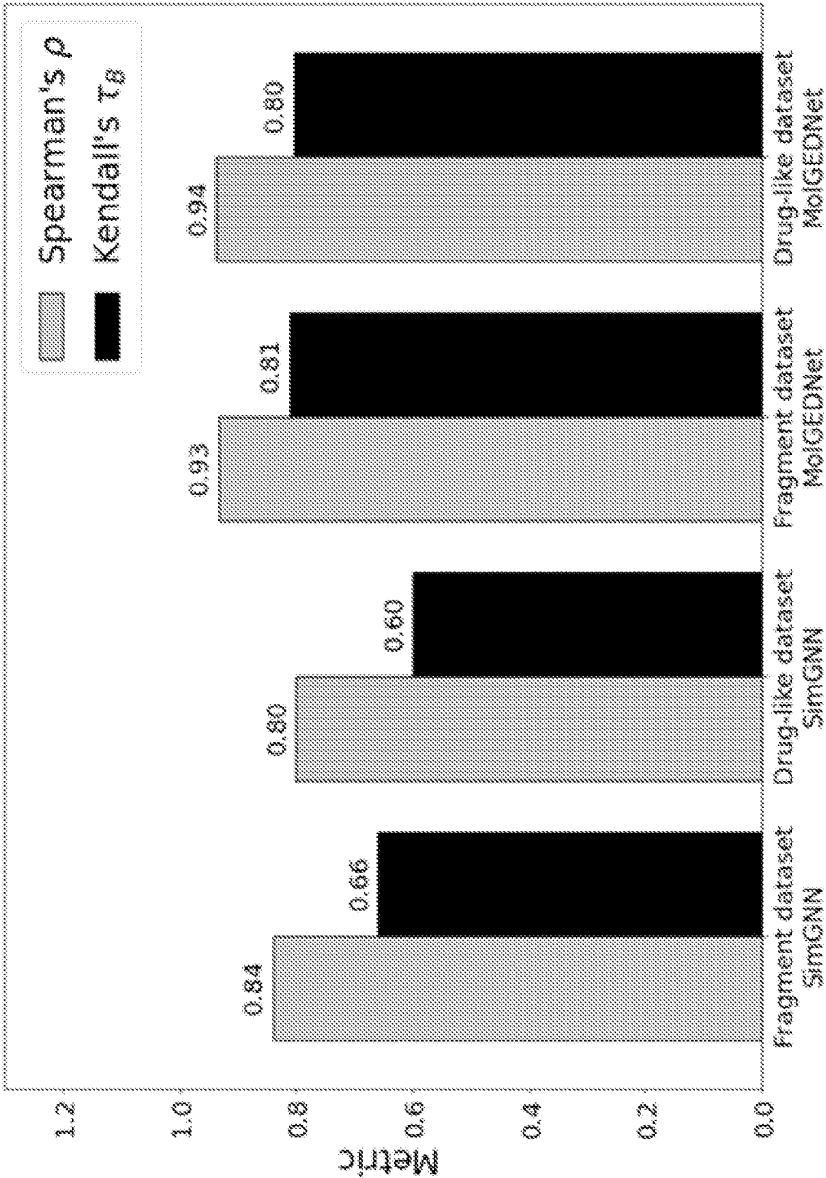
MolGEDNet significantly outperforms SimGNN on all metrics on both datasets



- MolGEDNet reduces MSE by around 60% (lower is better)

Fig. 8

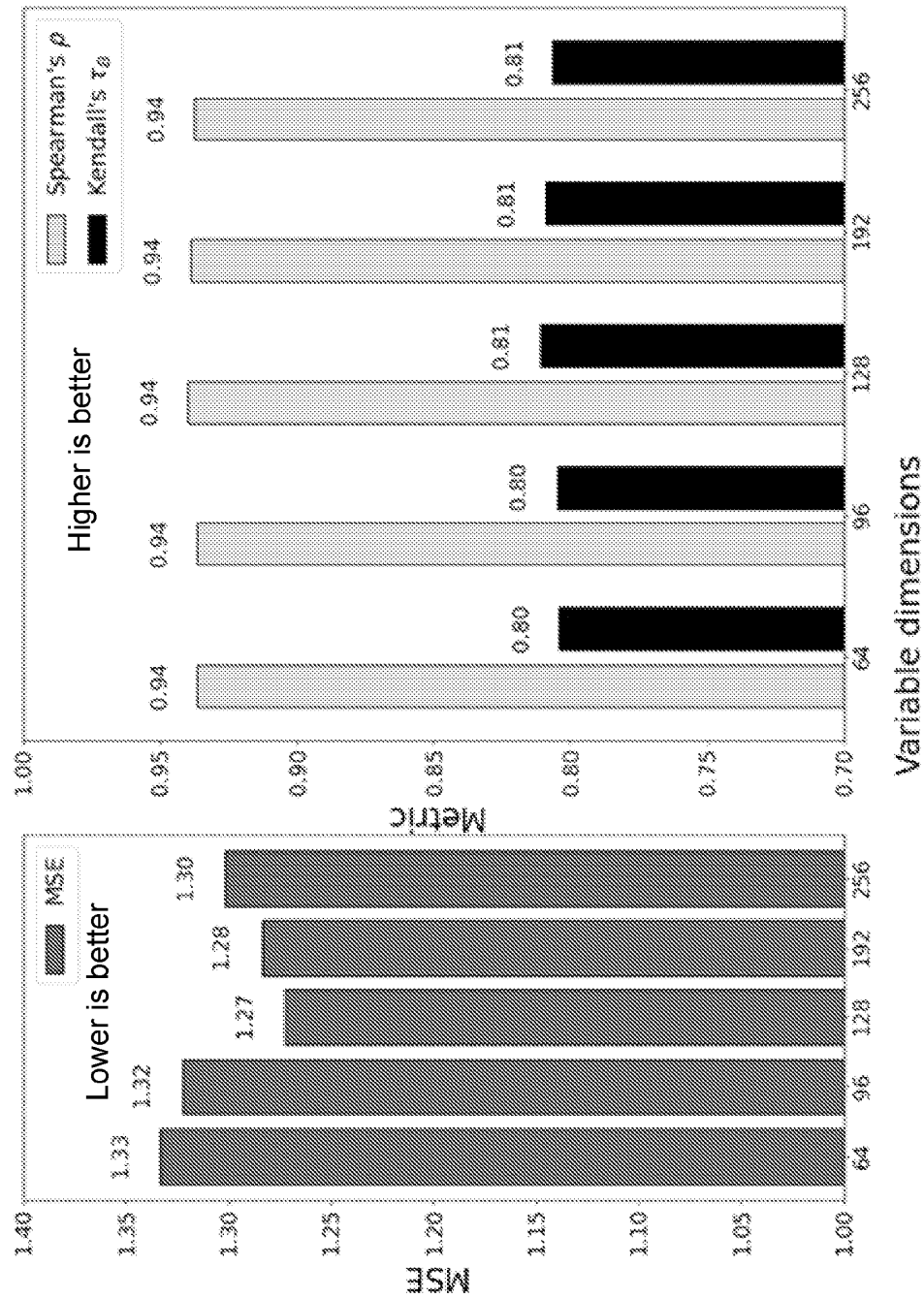
Performance of MoIGEDNet
MoIGEDNet significantly outperforms SimGNN on all metrics on both datasets



- MoIGEDNet increases τ_β by 22% to 35% (higher is better)

Fig. 9

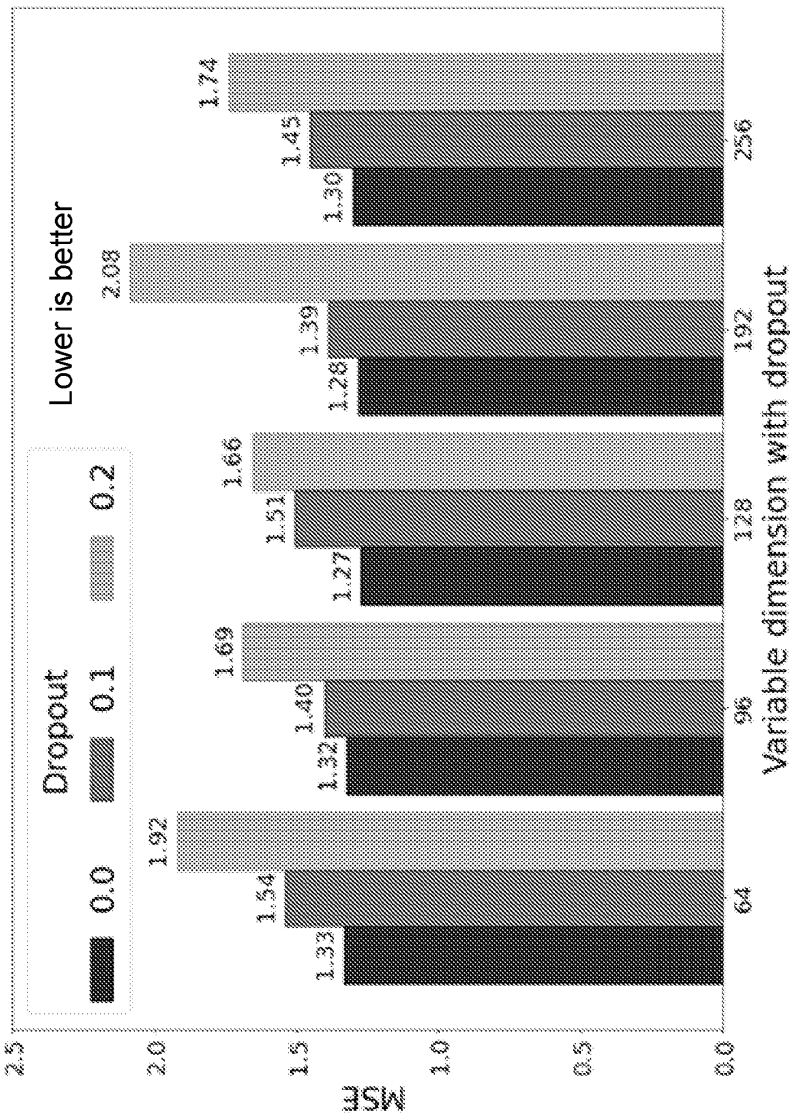
Hyperparameter tuning - results shown for drug-like dataset



- Hyperparameter tuning on drug-like dataset
- Increasing the model dimension does not impact model performance

Fig. 10

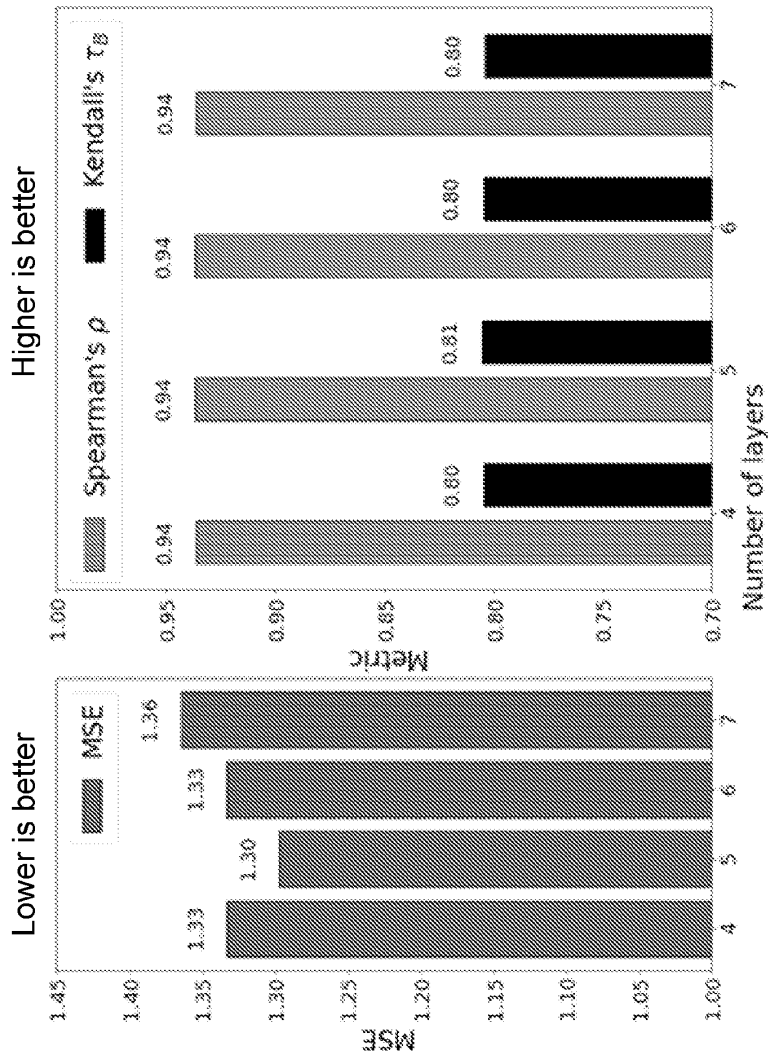
Hyperparameter tuning - results shown for drug-like dataset



- Hyperparameter tuning on drug-like dataset
- Increasing the model dimension does not impact model performance
- Train MSE < Test MSE
- Addition of dropout deteriorates performance

Fig. 11

Hyperparameter tuning - results shown for drug-like dataset



- Hyperparameter tuning on drug-like dataset
- Increasing the model dimension does not impact model performance
- Train MSE < Test MSE
- Addition of dropout deteriorates performance
- Increasing number of layers to 5 improves performance

Fig. 12

Hyperparameter tuning

Putting the best set of hyperparameters together

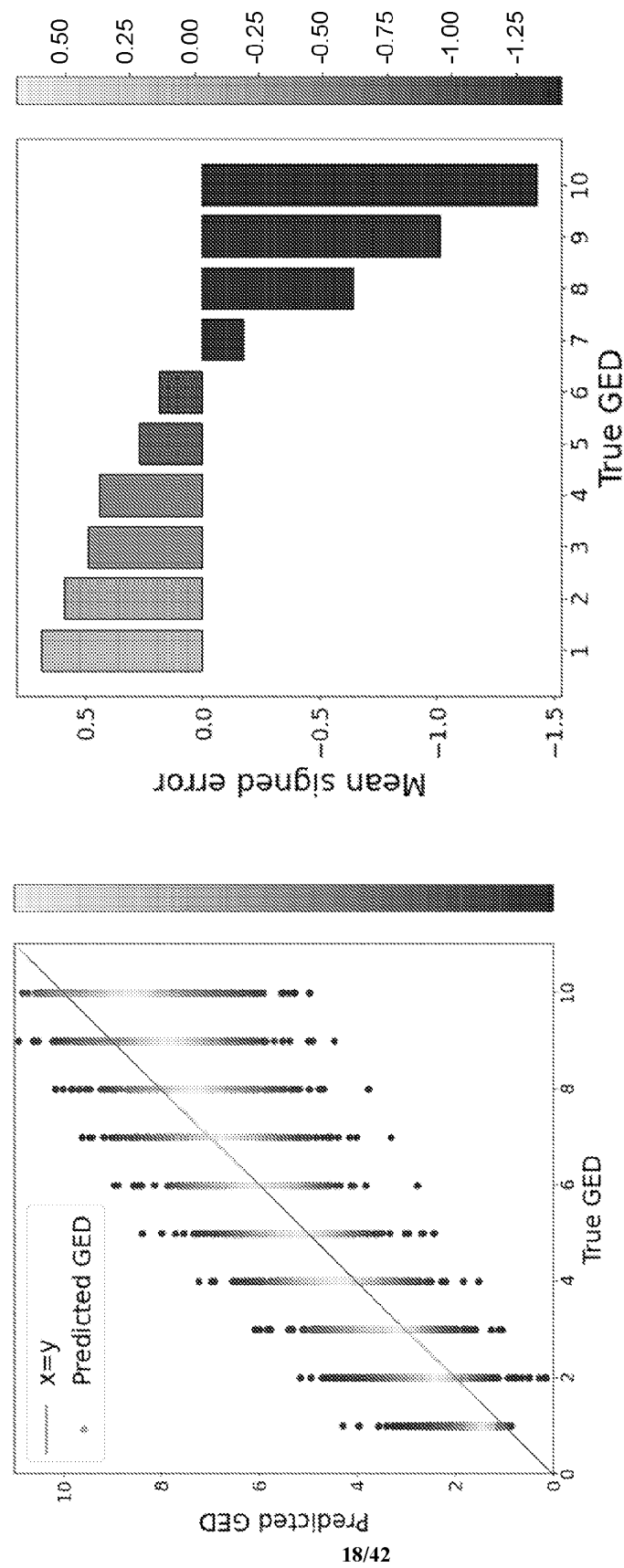
Best set of hyperparameters			
Variable dimensions	Number of layers	Dropout	
128	5	0.0	

MSE	Spearman's rho	Kendall's tau_b
1.27	0.94	0.81

- With the best set of hyperparameters, MSE on test set is 1.27
- Ranking correlation metrics remain more or less the same

Fig. 13

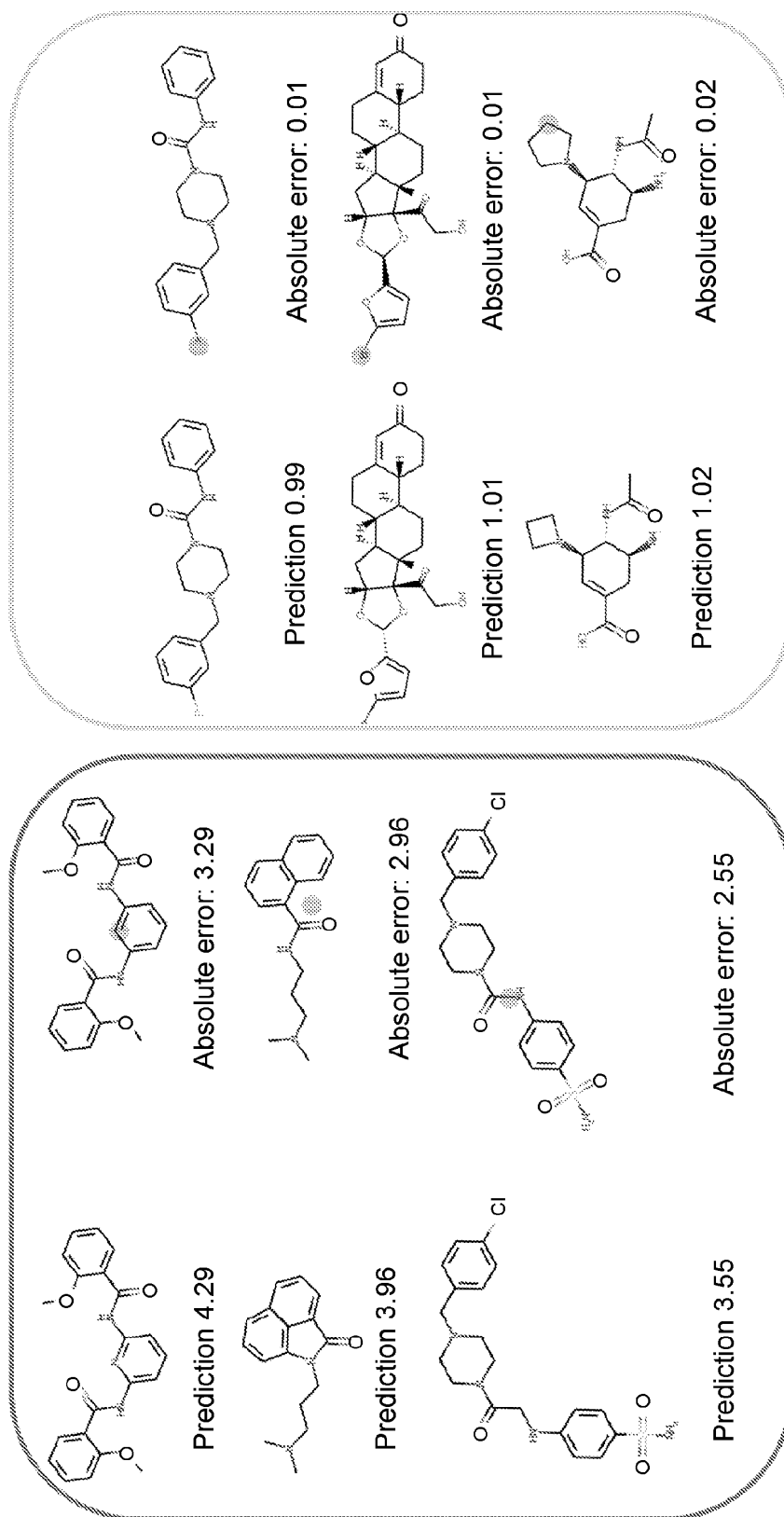
Error Analysis



- MolGEDNet predictions have mean predictor tendency
- Underestimation on the high GED end is more pronounced than at low GED

Fig. 14

Error Analysis

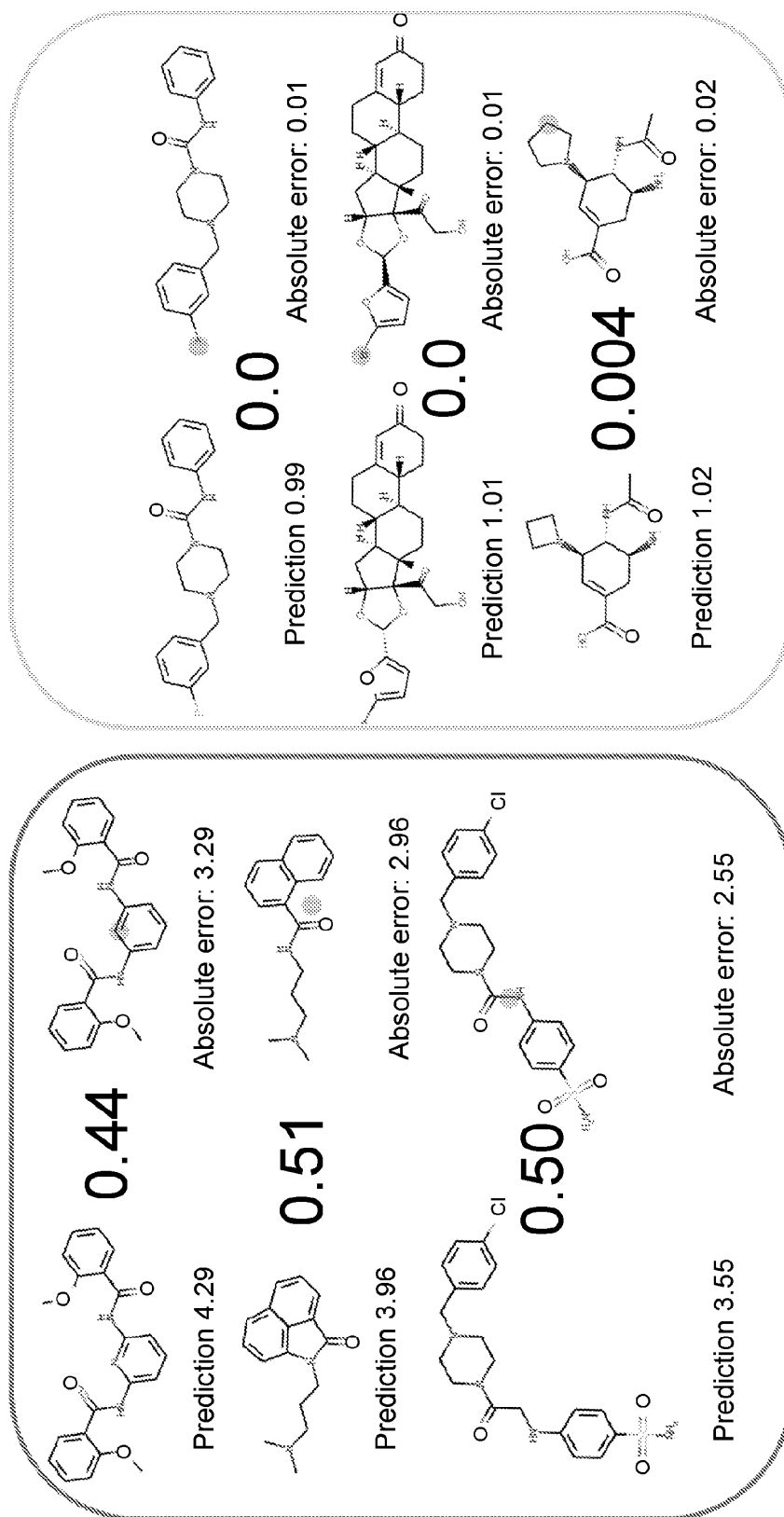


Large errors when edits are in the center of the molecule

High accuracy when edits are in the periphery of the molecule

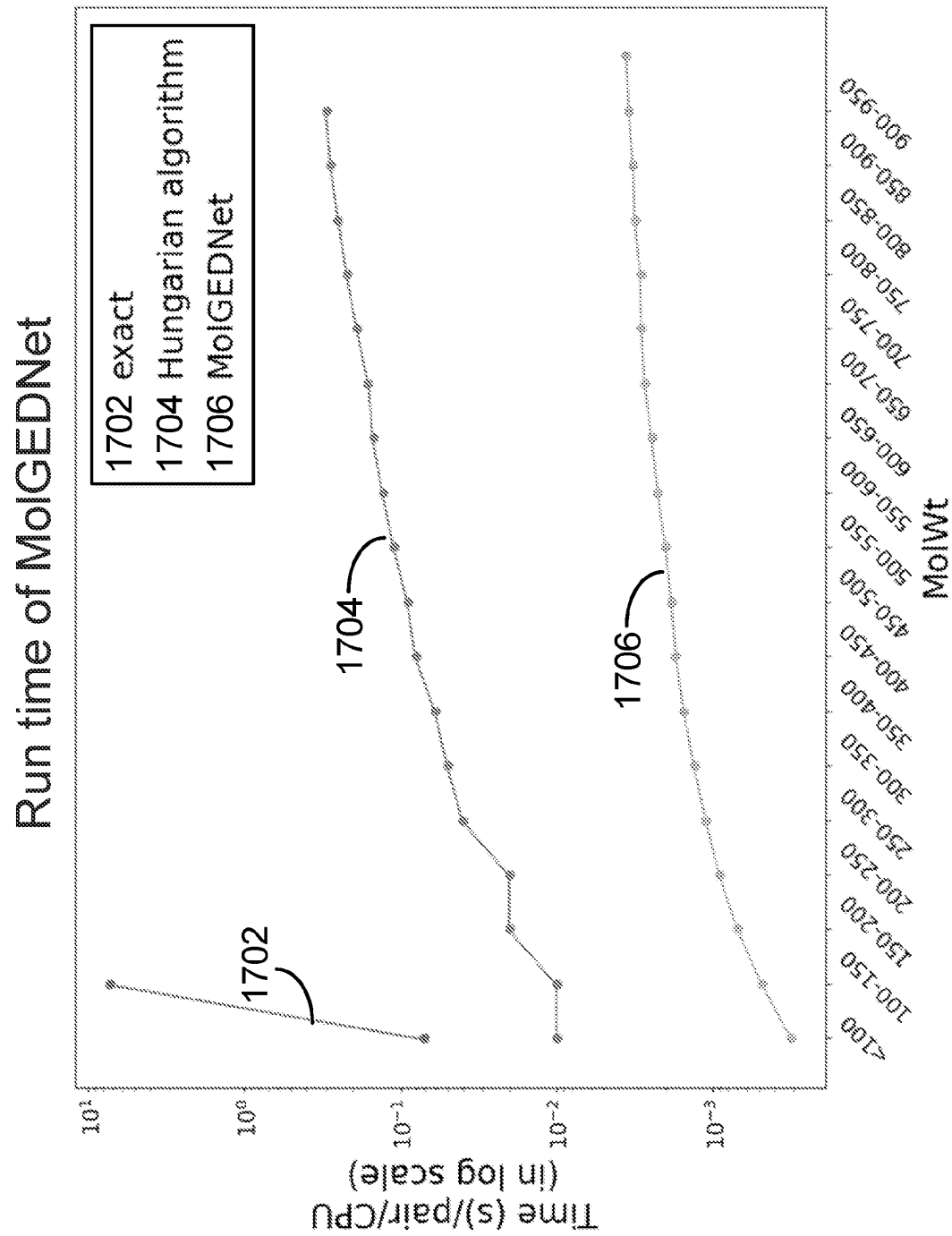
Fig. 15

Error Analysis



High error when edits are at large betweenness centrality nodes

Fig. 16



- Run time of MolGEDNet scales with the molecular weight (or number of atoms)
- MolGEDNet is ~50 times faster than the Hungarian algorithm for drug-like sized molecules

Fig. 17

Proposal for new architecture (e.g., MoIGEDNet)

Enforcing an isometry between GED metric and latent space

- * Train an encoder f that maps a graph G to a latent space $h \in \mathbb{R}^D$

$$h_i = f(G_i)$$

$$s.t. |h_i - h_j|_2 = GED(G_i, G_j)$$
- * Properties of a metric are satisfied by construction
- * Use more expressive graph neural networks like message passing neural networks for encoder f

Fig. 18

MoIGEDNet

Node, edge, and graph updates in MoIGEDNet

MPNN updates node, edge, and graph features as follows

$$\Delta \mathbf{e}_{ij}^{l+1} = \text{EdgeModel}(\mathbf{e}_{ij}^l, \mathbf{x}_i^l, \mathbf{x}_j^l, \mathbf{h}^l)$$
$$\Delta \mathbf{x}_i^{l+1} = \text{NodeModel}(\sum_{j \in \mathcal{N}(i)} \mathbf{e}_{ij}^{l+1}, \mathbf{x}_i^l, \mathbf{h}^l)$$
$$\Delta \mathbf{h}^{l+1} = \sum_{i \in \mathcal{V}} \mathbf{x}_i^{l+1}$$

Variables

$\mathbf{x}_i^l \in \mathbb{R}^{64}$

$\mathbf{e}_{ij}^l \in \mathbb{R}^{64}$

$\mathbf{h}^l \in \mathbb{R}^{64}$

$l = 4$

Fig. 19

Message Passing Neural Network details

$$G = (\mathcal{V}, \mathcal{E})$$

$$\{\mathbf{x}_i^0\} \quad i \in \mathcal{V}$$

$$\{\mathbf{e}_{ij}^0\} \quad (i, j) \in \mathcal{E}$$

$$\Delta \mathbf{x}_i^{l+1} = \text{Linear}(\text{ReLU}(\text{Linear}(\text{ReLU}(\text{LayerNorm}([\mathbf{x}_i^l | \mathbf{m}_i^{l+1} | \mathbf{g}^l])))$$

$$\Delta \mathbf{e}_{ij}^{l+1} = \text{Linear}(\text{ReLU}(\text{Linear}(\text{ReLU}(\text{LayerNorm}([\mathbf{e}_{ij}^l | \mathbf{g}^l | \mathbf{x}_i^l | \mathbf{x}_j^l])))$$

$$\Delta \mathbf{g}^{l+1} = \sum_{i \in \mathcal{V}} \mathbf{n}_i^{l+1}$$

Fig. 20

Exact GED versus Chemical GED

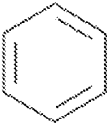
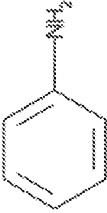
Exact GED		Chemical GED	
			1 • 1 addition of -NH2 group
			4 • 1 C - C bond deletion • 1 atom addition • 2 C - N bond additions 1 • 1 addition of NH group

Fig. 21
(Prior Art)

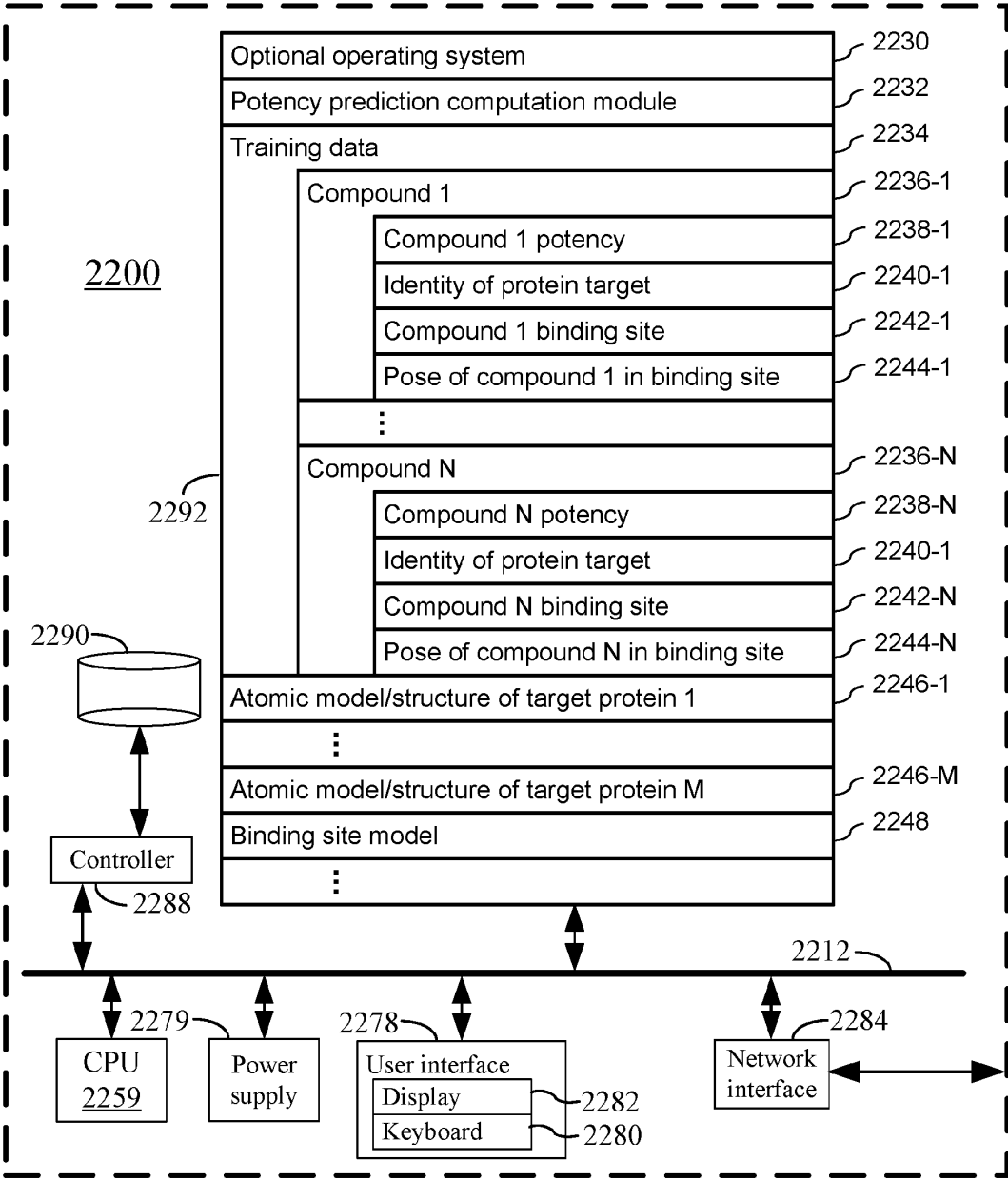


Fig. 22

(2300) System and methods for training a model to predict potency of a test compound against an orthosteric site of a target protein are provided.



(2302) Obtain a dataset comprising data for a plurality of training compounds. Each training compound is labeled with a respective protein target in a plurality of protein targets, a binding site for the respective protein target and a potency against the protein target.

(2304) Each training compound satisfies two or more rules, three or more rules, or all four rules of the Lipinski's rule of Five: (i) not more than five hydrogen bond donors, (ii) not more than ten hydrogen bond acceptors, (iii) a molecular weight under 500 Daltons, and (iv) a LogP under 5.

(2306) Each training compound is an organic compound having a molecular weight of less than 500 Daltons, less than 1000 Daltons, less than 2000 Daltons, less than 4000 Daltons, less than 6000 Daltons, less than 8000 Daltons, less than 10000 Daltons, or less than 20000 Daltons.

(2308) Each training compound is an organic compound having a molecular weight of between 400 Daltons and 10000 Daltons.

(2310) The plurality of training compounds comprises 1000 or more, 2000 or more or 10,000 or more training compounds.

(2312) Each of the protein targets is represented by one or more compounds in the plurality of compounds and the plurality of protein targets comprises 10 or more protein targets or 100 or more protein targets.

(2314) At least 5 percent, 10 percent, or 15 percent of the protein targets comprise multiple binding sites.



A

Fig. 23A

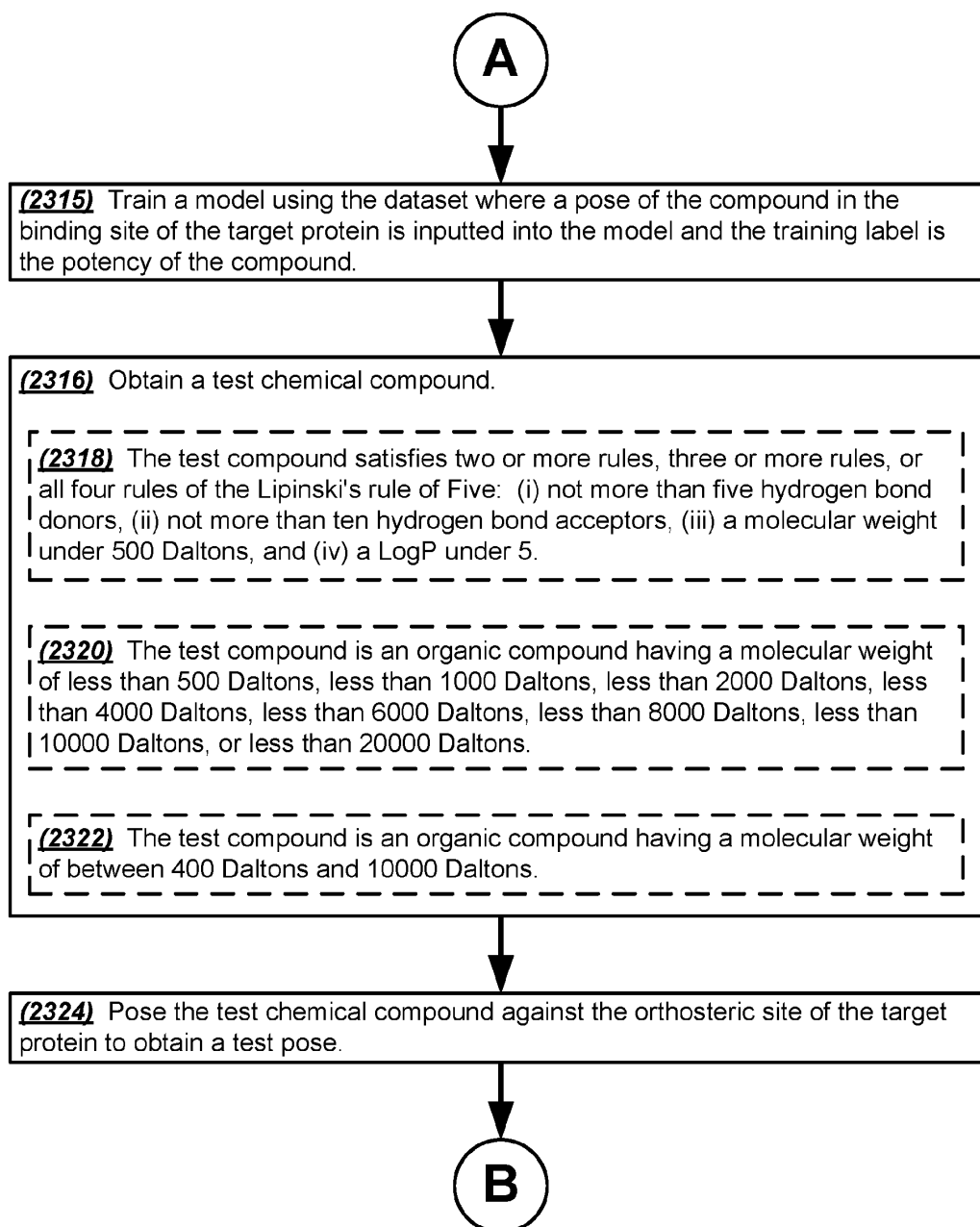


Fig. 23B

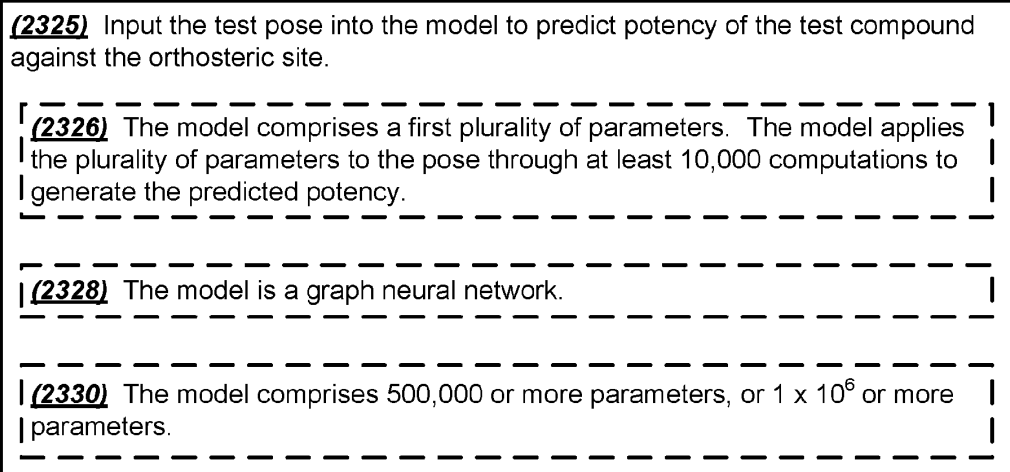
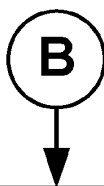


Fig. 23C

Compound Selection Workflow

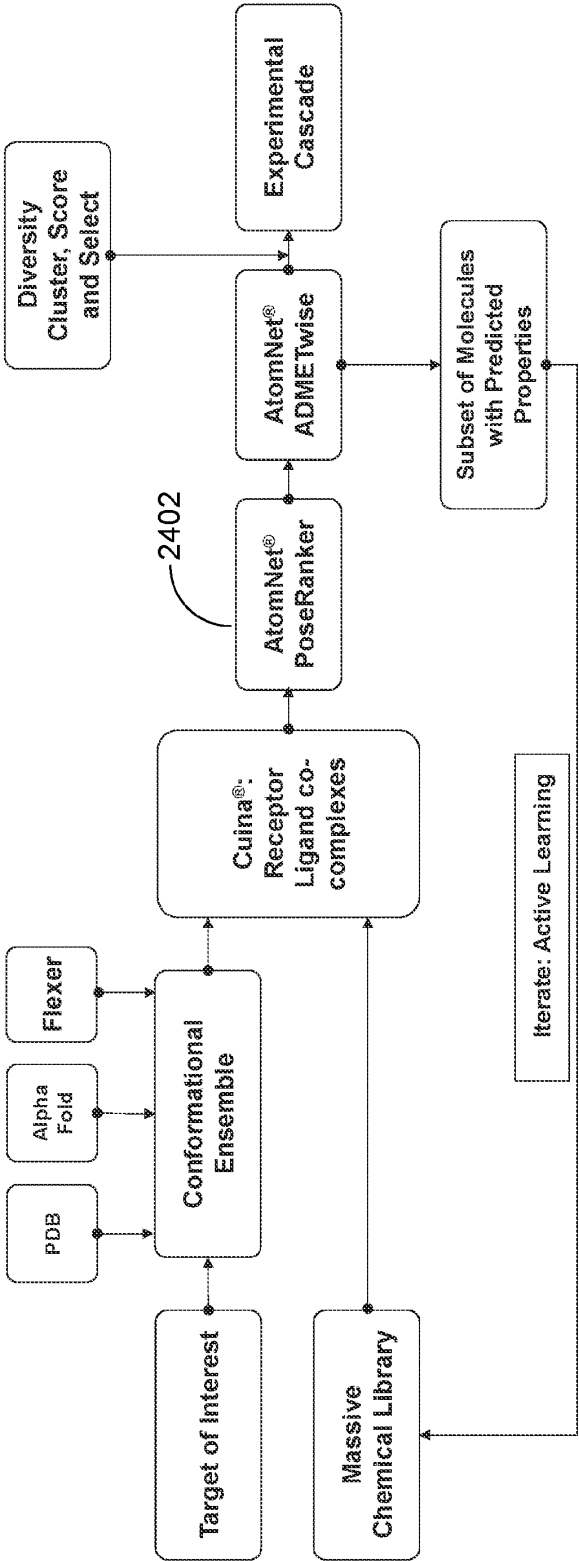


Fig. 24

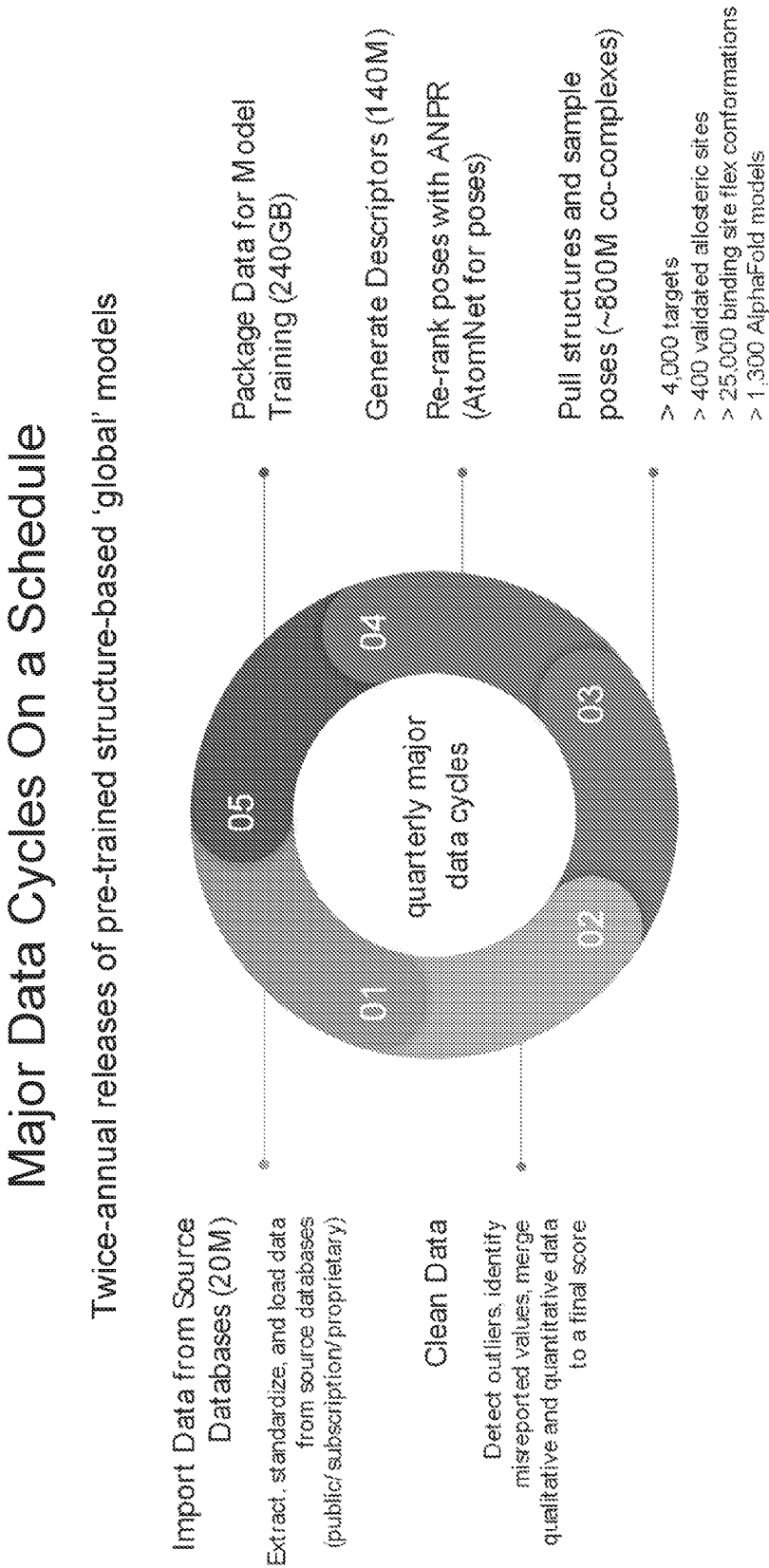


Fig. 25

Benchmarks: Do Our Models Generalize?

Stringent holdout sets test predictive power on multiple facets of early discovery

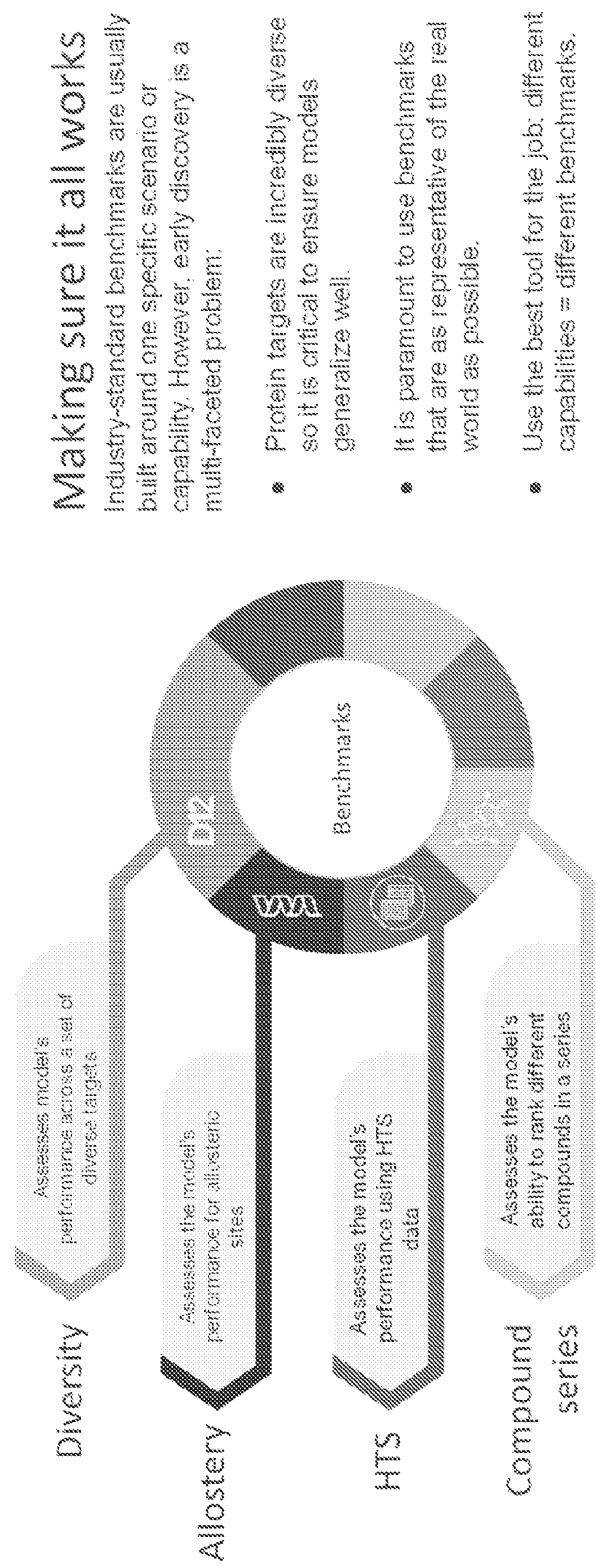


Fig. 26

Anti-Benchmarks: do our models cheat?

Our anti-benchmarks guarantee that our models are held to the highest standards

Opening the DL black box

We carefully developed state-of-the-art anti-benchmarks to gain insights into the behavior of our models.

- Are our models learning physics?
- Are our models focusing on the correct features of the protein?
- Identifying cheating: are the models just learning how to pass the test?

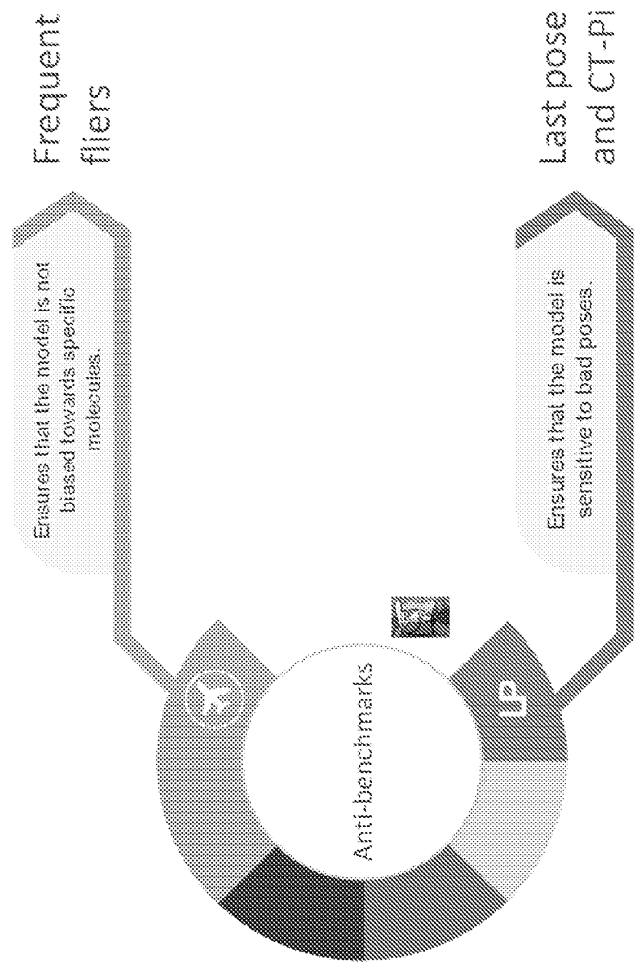


Fig. 27

AlphaFold2 helps structure-based predictions

AF2 structurally enables underrepresented, therapeutically relevant protein classes

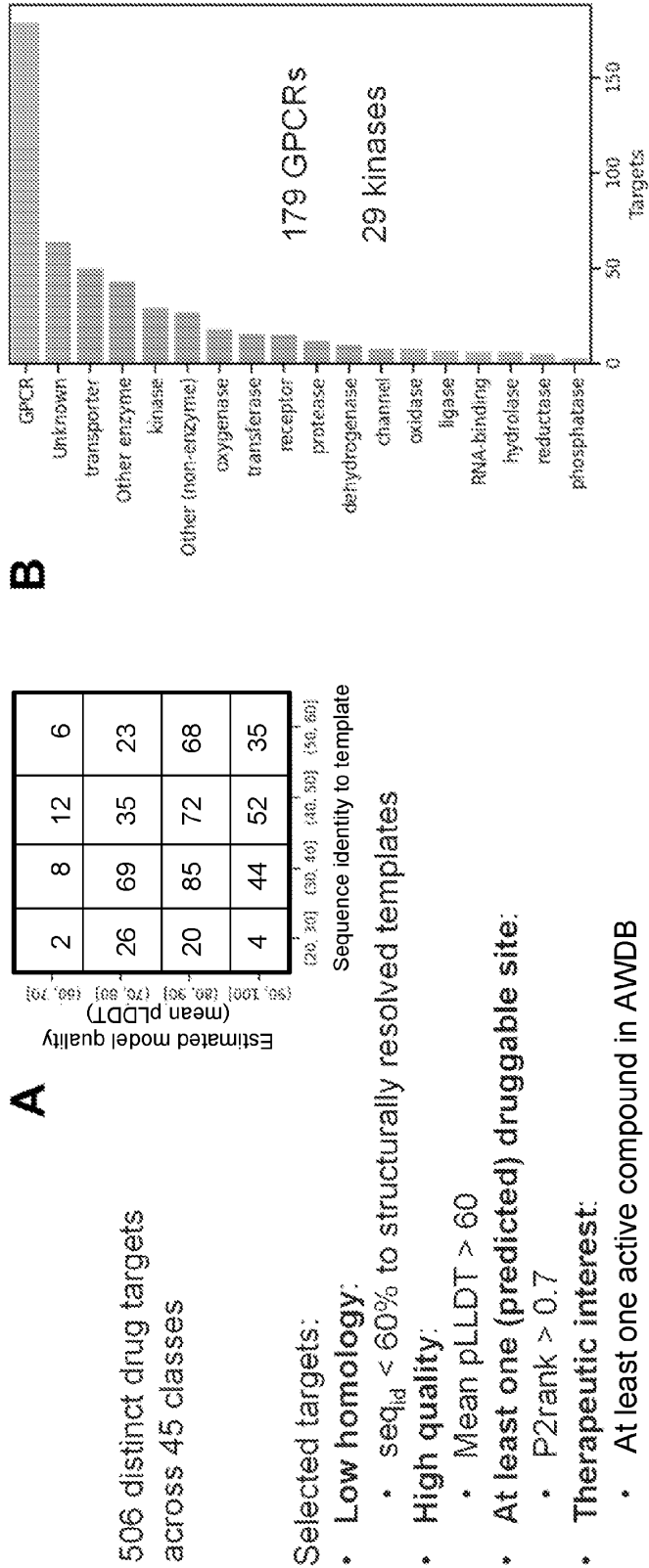


Fig. 28

Where do active compounds bind?

Most activity measurements cannot be reliably mapped to a specific binding site

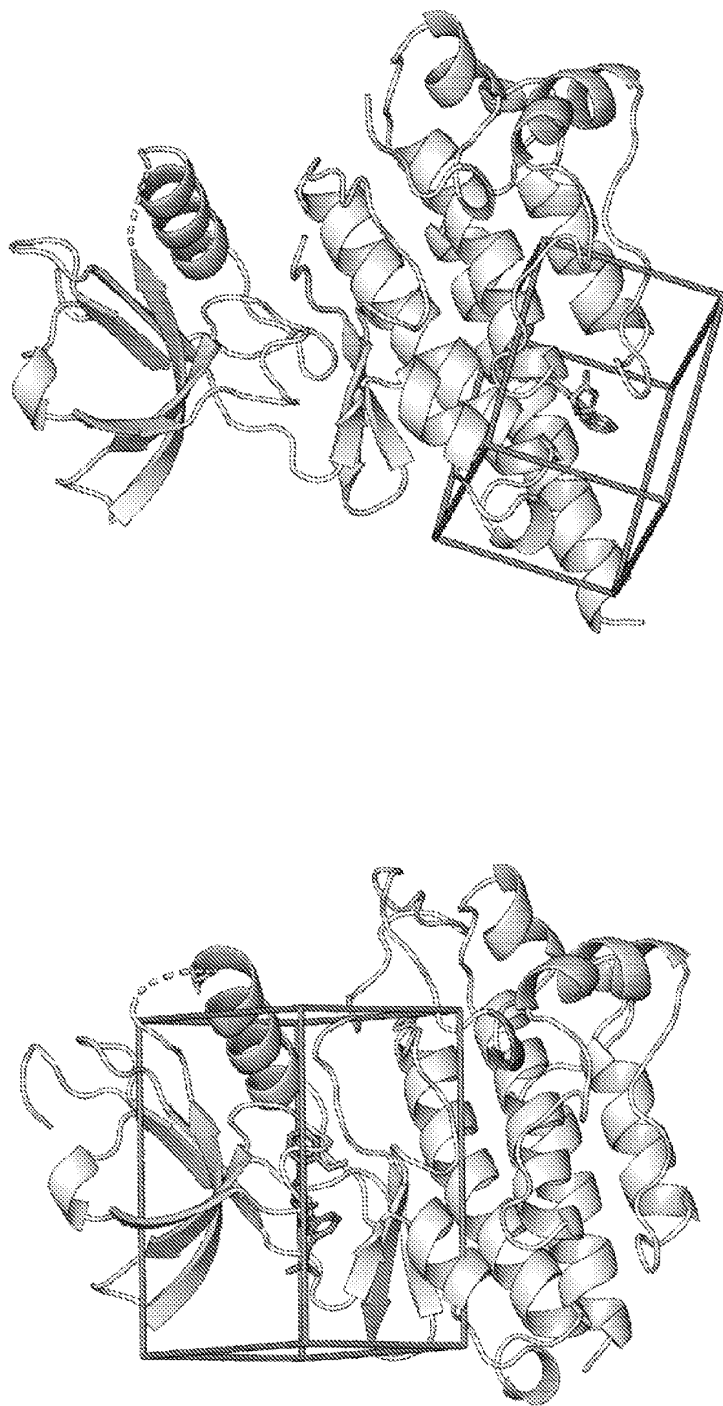


Fig. 29A

Mapping Known Actives to Detected Sites

Unique approach to generate large databases

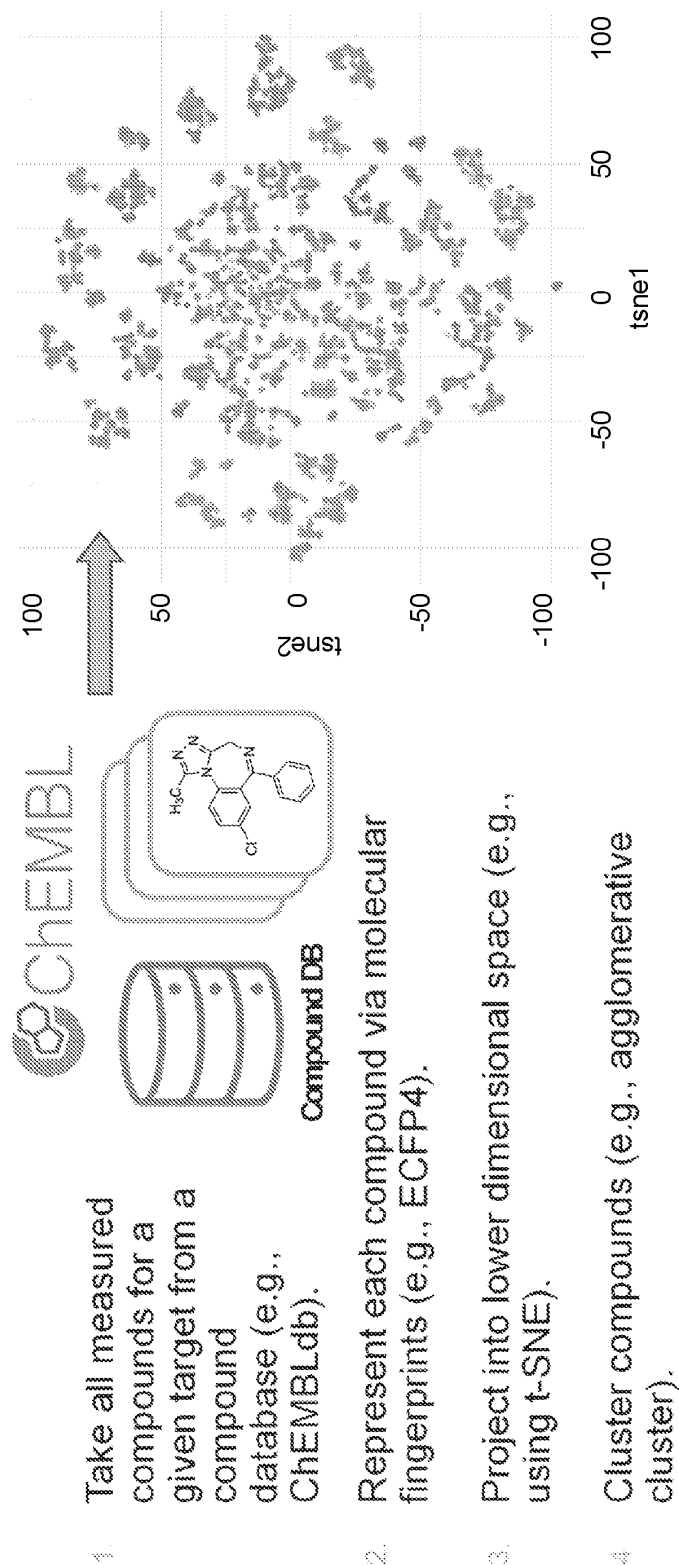


Fig. 29B

Mapping Known Actives to Detected Sites

We can use the compound clusters + structural data to map more compounds

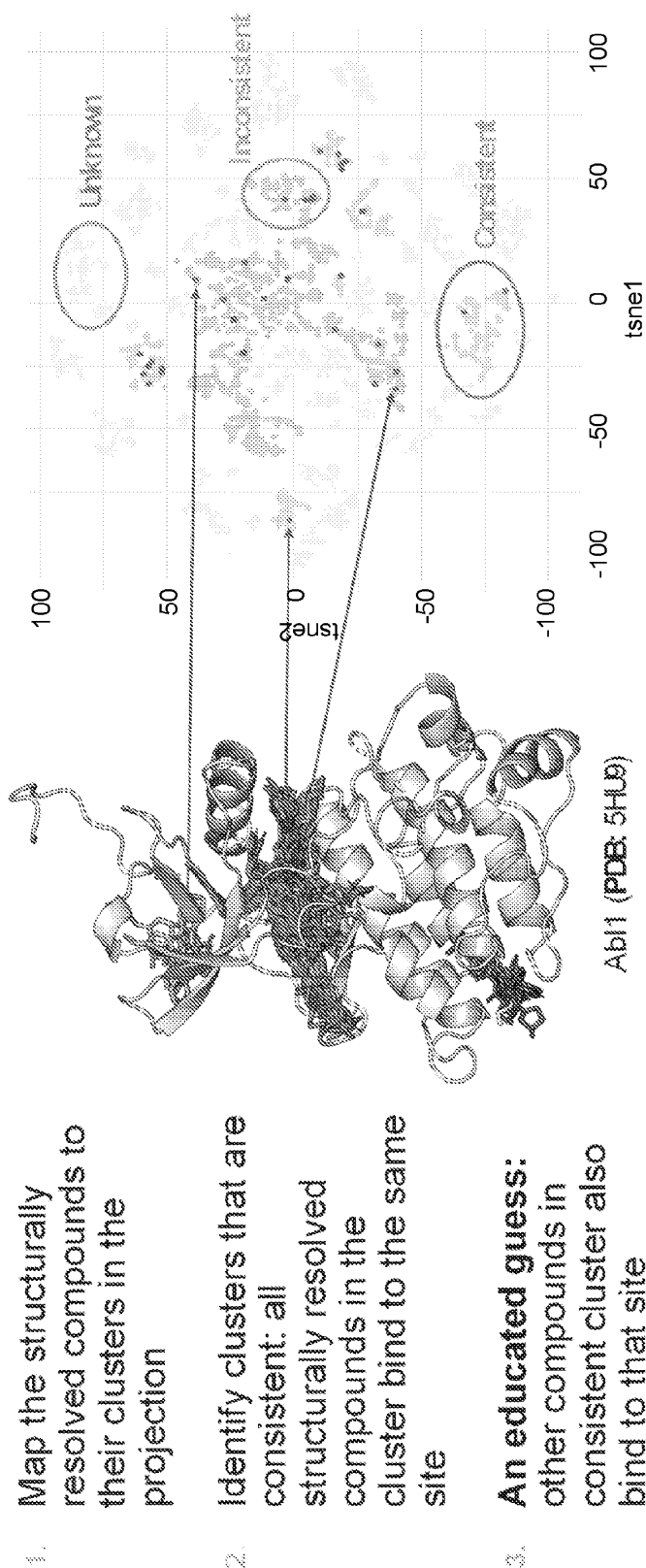


Fig. 30

Mapping bioactives to structurally dark targets

AF2 structurally enables underrepresented, therapeutically relevant protein classes

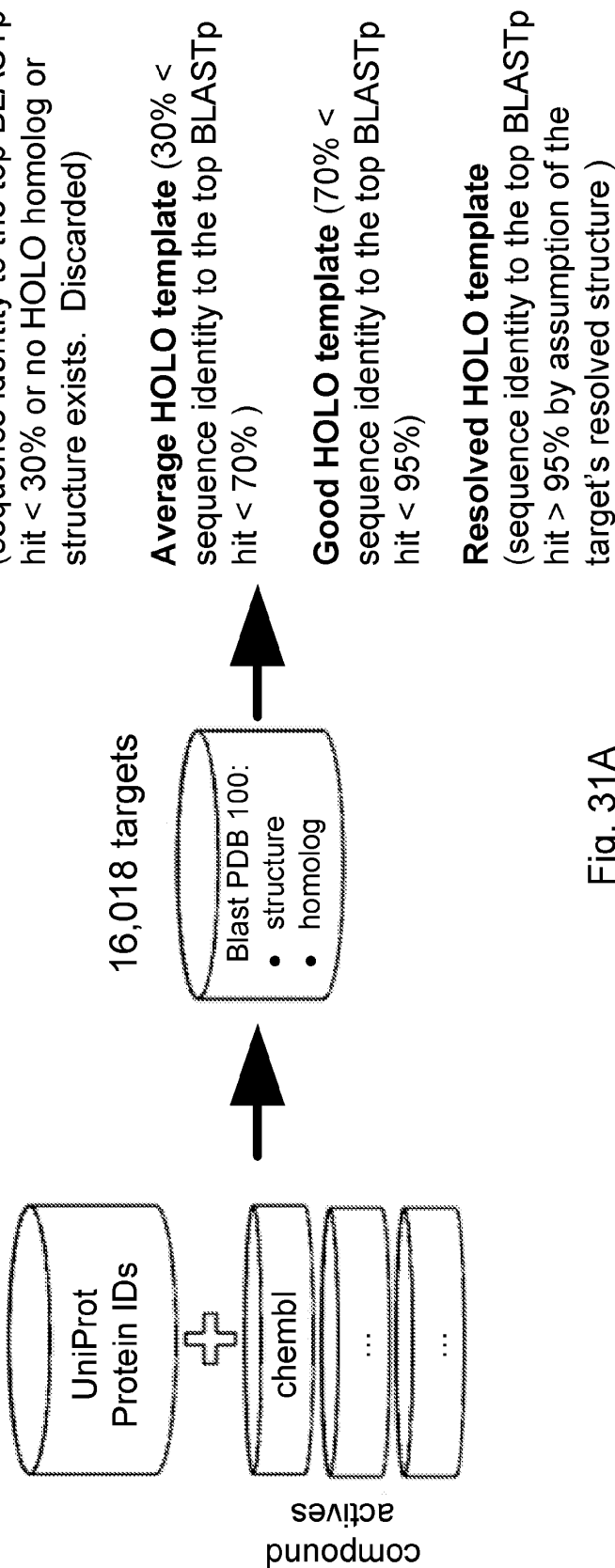


Fig. 31A

Mapping bioactives to structurally dark targets

AF2 structurally enables underrepresented, therapeutically relevant protein classes

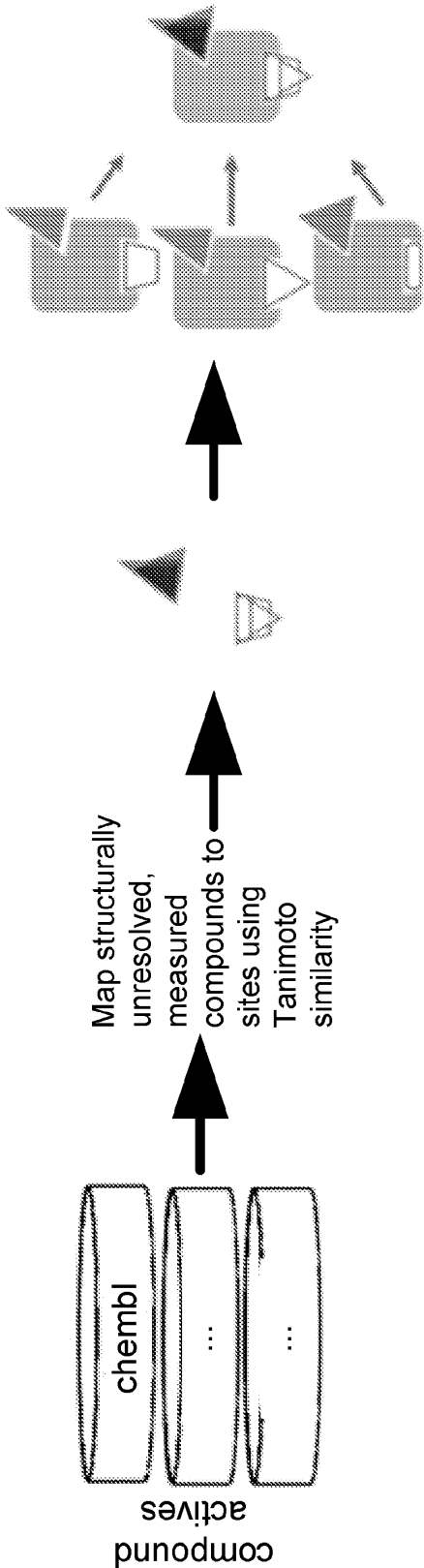


Fig. 31B

Multi-Site Labeling Deconvolves the SAR Data

Clean multi-site data improves vHTS performance – even on primary site!

- These models were trained using data from public databases such as ChEMBLdb.
- No binding sites for the proteins in Atomwise test set were used during training (70% sequence similarity split).
- Improvement in performance for primary site highlights the impact of incorrect data labeling.

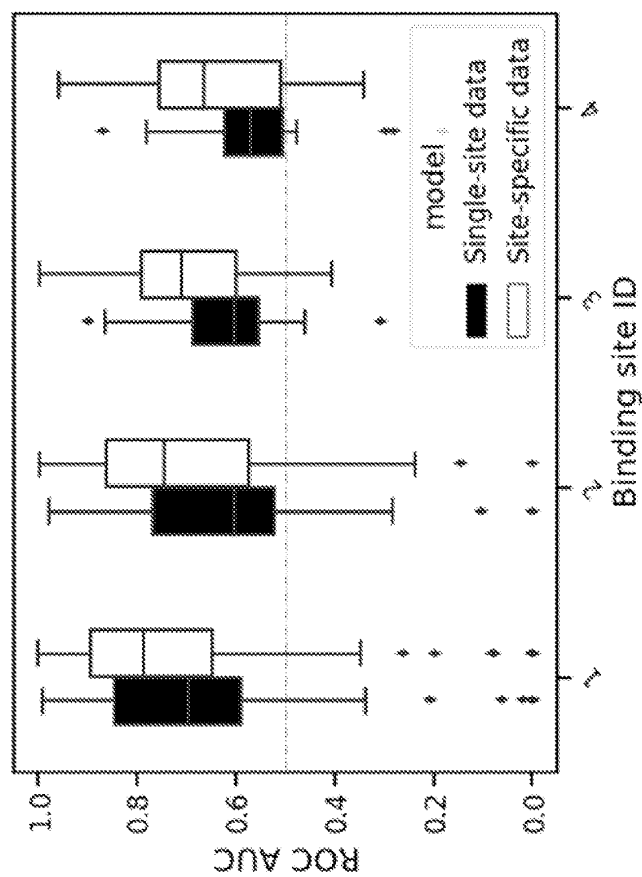


Fig. 32

GPCR-specific AtomNet model

Training with AF2 structures improves potency prediction on GLASS holdout

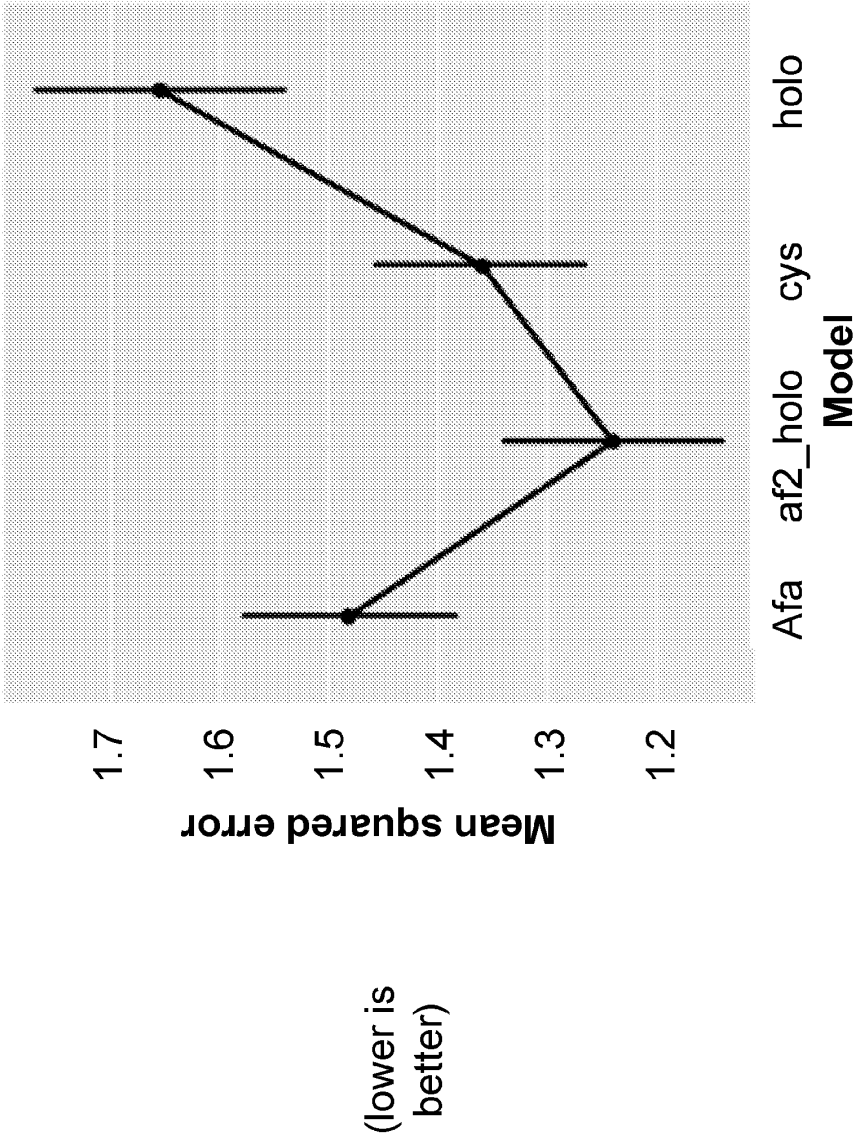


Fig. 33A

GPCR-specific AtomNet model

Training with AF2 structures improves potency prediction on GLASS holdout

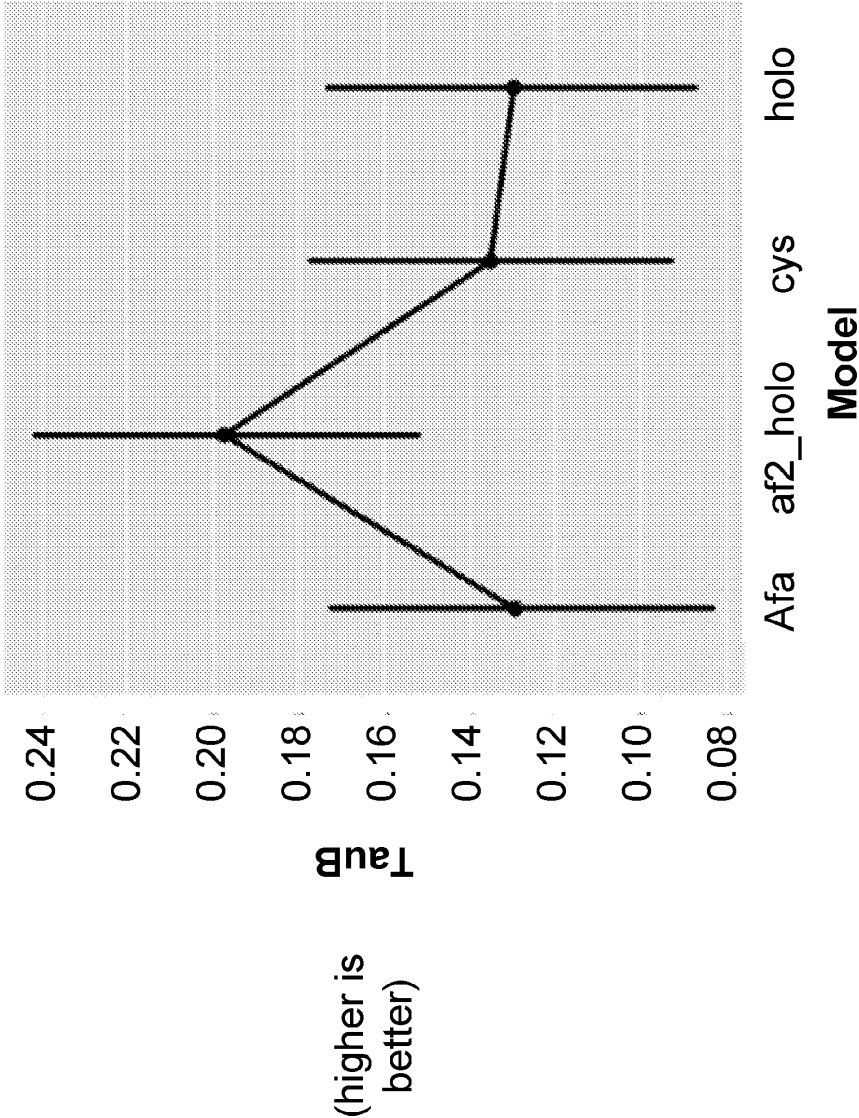


Fig. 33B