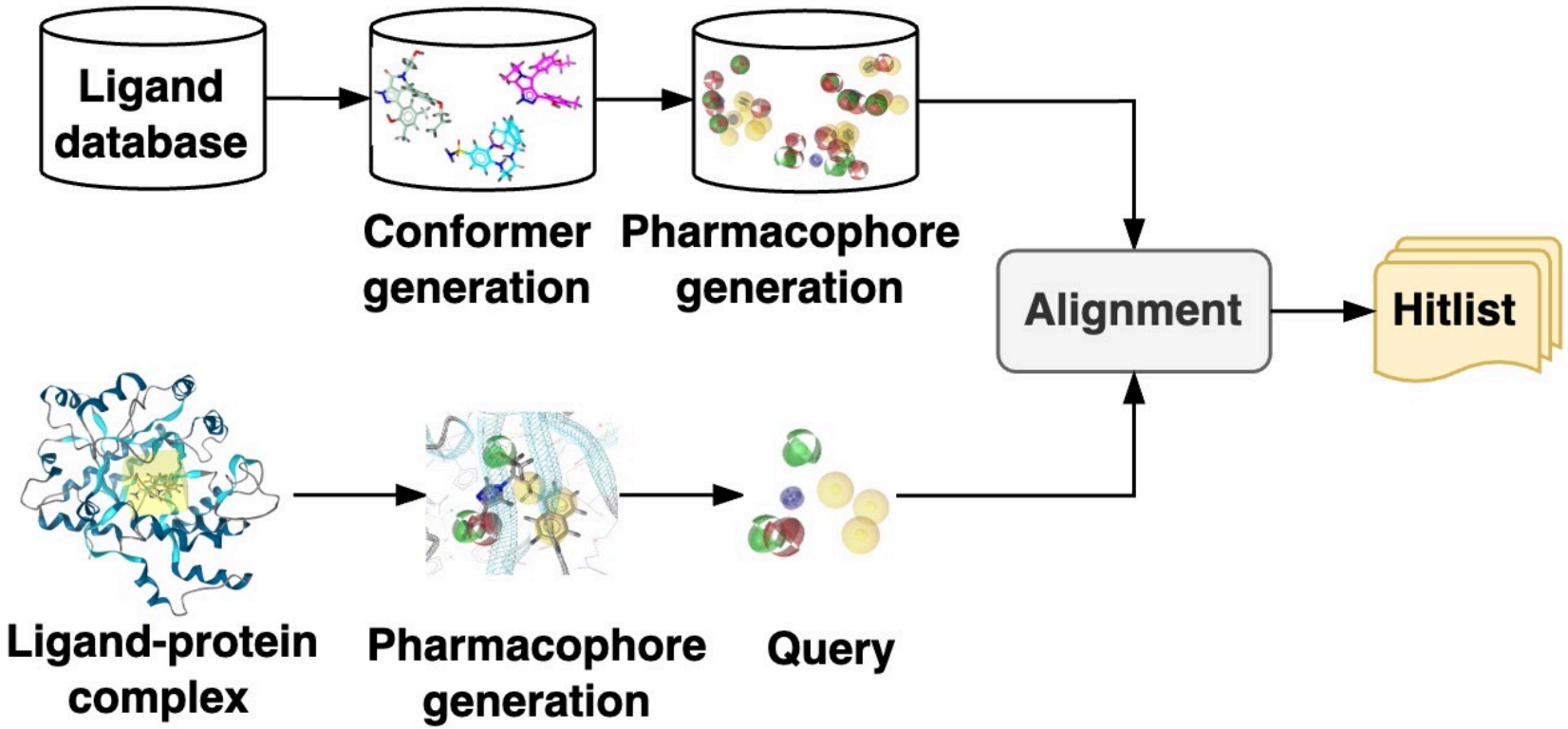
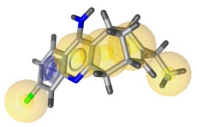
PharmacoMatch: Efficient 3D Pharmacophore Screening via Neural Subgraph Matching

Daniel Rose[1,2,3], Oliver Wieder[1,2] , Thomas Seidel[1,2], Thierry Langer[1,2]

[1]Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria
[2]Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Josef-Holaubek-Platz 2, 1090 Vienna, Austria
[3]Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria

Pharmacophore Screening





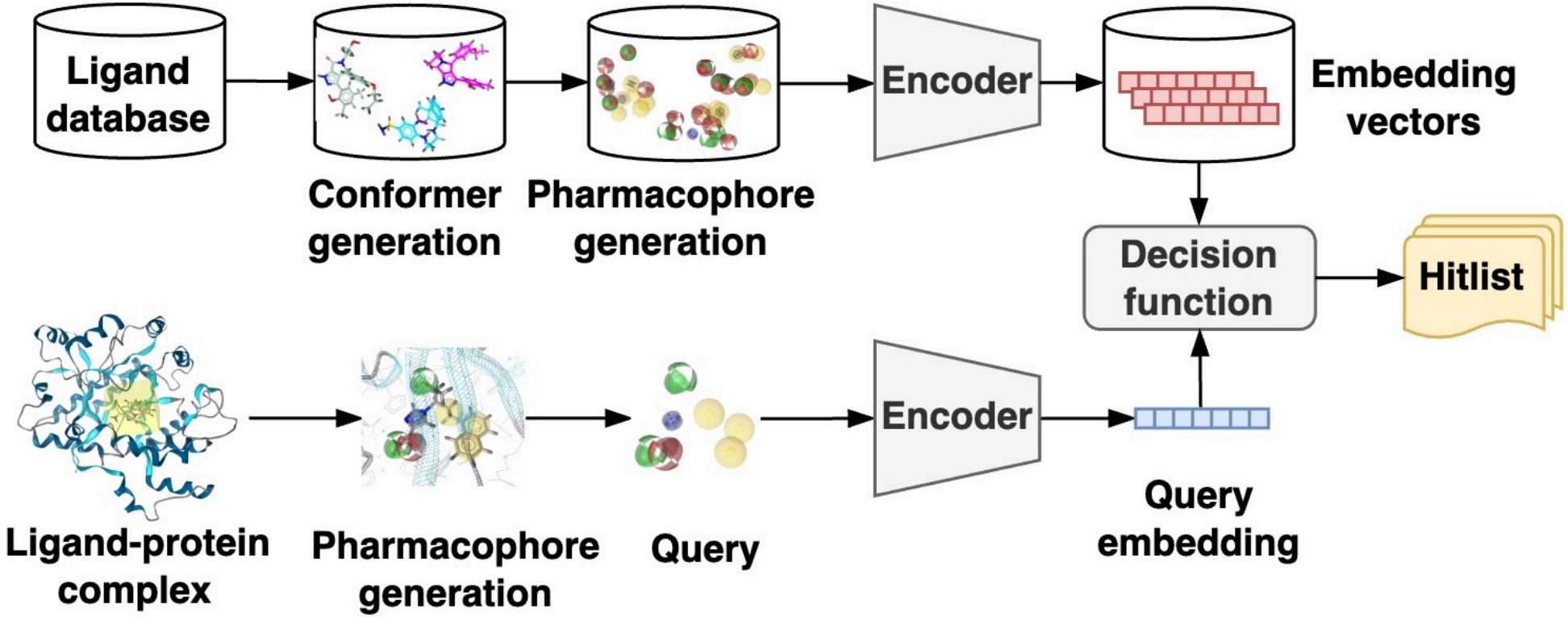
3D Pharmacophore screening works with alignment algorithm

😔

Alignment does not scale well with large molecular databases

Can we train a model to speed up the alignment?

Overview: PharmacoMatch



- Conceptualization of pharmacophore matching as representation learning problem
- Model training via self-supervised learning strategy
- Virtual pre-screening with learned embeddings is efficient and scalable

Pharmacophore Matching

Pharmacophore representation

$$P = \{(\mathbf{r}_i, d_i) \in \mathbb{R}^3 \times \mathcal{D}\}_i$$

- Pharmacophore P
- Cartesian coordinates $\mathbf{r}_i$
- Descriptor d_i
- Descriptor set $\mathcal{D}$

$$G(P) = (V_P, E_P, \lambda_P)$$

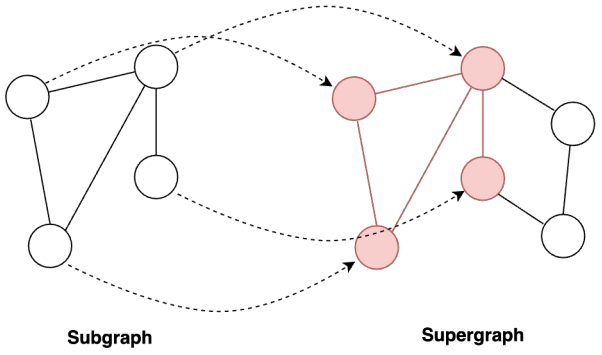
- Complete graph G
- Node set $V_P = \{v_1, ..., v_{|P|}\}$
- Edge set $E_P = V_P \times V_P$
- Label set $\mathcal{L} = \mathcal{D} \cup \mathcal{R}$
- $\mathcal{R} = \{\|\mathbf{r}_i - \mathbf{r}_j\|_2 \mid 1 \leq i, j \leq |P|\}$
- Labeling function $\lambda_P : V \cup E \to \mathcal{L}$
- Node attributes $\lambda_P(v_i) = l_i$
- Edge attributes $\lambda_P(e_{ij}) = \|\mathbf{r}_i - \mathbf{r}_j\|_2$

Subgraph isomorphism

$$G_1 = (V_1, E_1, \lambda_1),\ G_2 = (V_2, E_2, \lambda_2)$$
$$G_1 \simeq G_2 \text{ iff } \exists f : V_1 \to V_2 \text{ s.t.}$$

- $\forall (u, v) \in E_1 : (f(u), f(v)) \in E_2$
- $\forall v \in V_1 : \lambda_1(v) = \lambda_2(f(v))$
- $\forall (u, v) \in E_1 : \lambda_1((u, v)) = \lambda_2((f(u), f(v)))$

$$G_Q = (V_Q, E_Q, \lambda_Q),\ G_T = (V_T, E_T, \lambda_T)$$
$$G_H = (V_H, E_H, \lambda_H) \text{ s.t. } V_H \subseteq V_T,\ E_H \subseteq E_T$$
$$\mathcal{H} = \{G_H \mid G_H \simeq G_Q\}$$
$$G_Q \lesssim G_T \text{ iff } \mathcal{H} \neq \varnothing$$

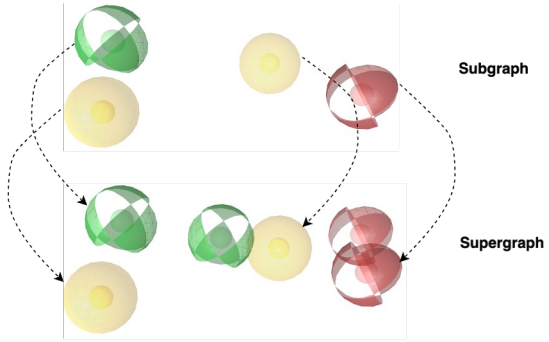


Pharmacophore matching

$$P_H \subseteq P_T \text{ iff } \exists g : P_Q \to P_H \text{ s.t.}$$

- $\forall i \in P_Q : d_i = d_{g(i)}$
- $\|\mathbf{r}_i - \mathbf{r}_{g(i)}\|_2 < r_T$
- r_T tolerance sphere radius

$$G_Q = G(P_Q),\ G_H = G(P_H),\ G_T = G(P_T)$$
$$\lambda_Q((u, v)) \approx \lambda_H((f(u), f(v)))$$



Pharmacophore Matching

Pharmacophore representation

$$P = \{(\mathbf{r}_i, d_i) \in \mathbb{R}^3 \times \mathcal{D}\}_i$$

- Pharmacophore P
- Cartesian coordinates $\mathbf{r}_i$
- Descriptor d_i
- Descriptor set $\mathcal{D}$

$$G(P) = (V_P, E_P, \lambda_P)$$

- Complete graph G
- Node set $V_P = \{v_1, ..., v_{|P|}\}$
- Edge set $E_P = V_P \times V_P$
- Label set $\mathcal{L} = \mathcal{D} \cup \mathcal{R}$
- $\mathcal{R} = \{\|\mathbf{r}_i - \mathbf{r}_j\|_2 \mid 1 \leq i, j \leq |P|\}$
- Labeling function $\lambda_P : V \cup E \to \mathcal{L}$
- Node attributes $\lambda_P(v_i) = l_i$
- Edge attributes $\lambda_P(e_{ij}) = \|\mathbf{r}_i - \mathbf{r}_j\|_2$

Subgraph isomorphism

$$G_1 = (V_1, E_1, \lambda_1),\ G_2 = (V_2, E_2, \lambda_2)$$
$$G_1 \simeq G_2 \text{ iff } \exists f : V_1 \to V_2 \text{ s.t.}$$

- $\forall (u, v) \in E_1 : (f(u), f(v)) \in E_2$
- $\forall v \in V_1 : \lambda_1(v) = \lambda_2(f(v))$
- $\forall (u, v) \in E_1 : \lambda_1((u, v)) = \lambda_2((f(u), f(v)))$

$$G_Q = (V_Q, E_Q, \lambda_Q),\ G_T = (V_T, E_T, \lambda_T)$$
$$G_H = (V_H, E_H, \lambda_H) \text{ s.t. } V_H \subseteq V_T,\ E_H \subseteq E_T$$
$$\mathcal{H} = \{G_H \mid G_H \simeq G_Q\}$$
$$G_Q \lesssim G_T \text{ iff } \mathcal{H} \neq \varnothing$$

Pharmacophore matching

$$P_H \subseteq P_T \text{ iff } \exists g : P_Q \to P_H \text{ s.t.}$$

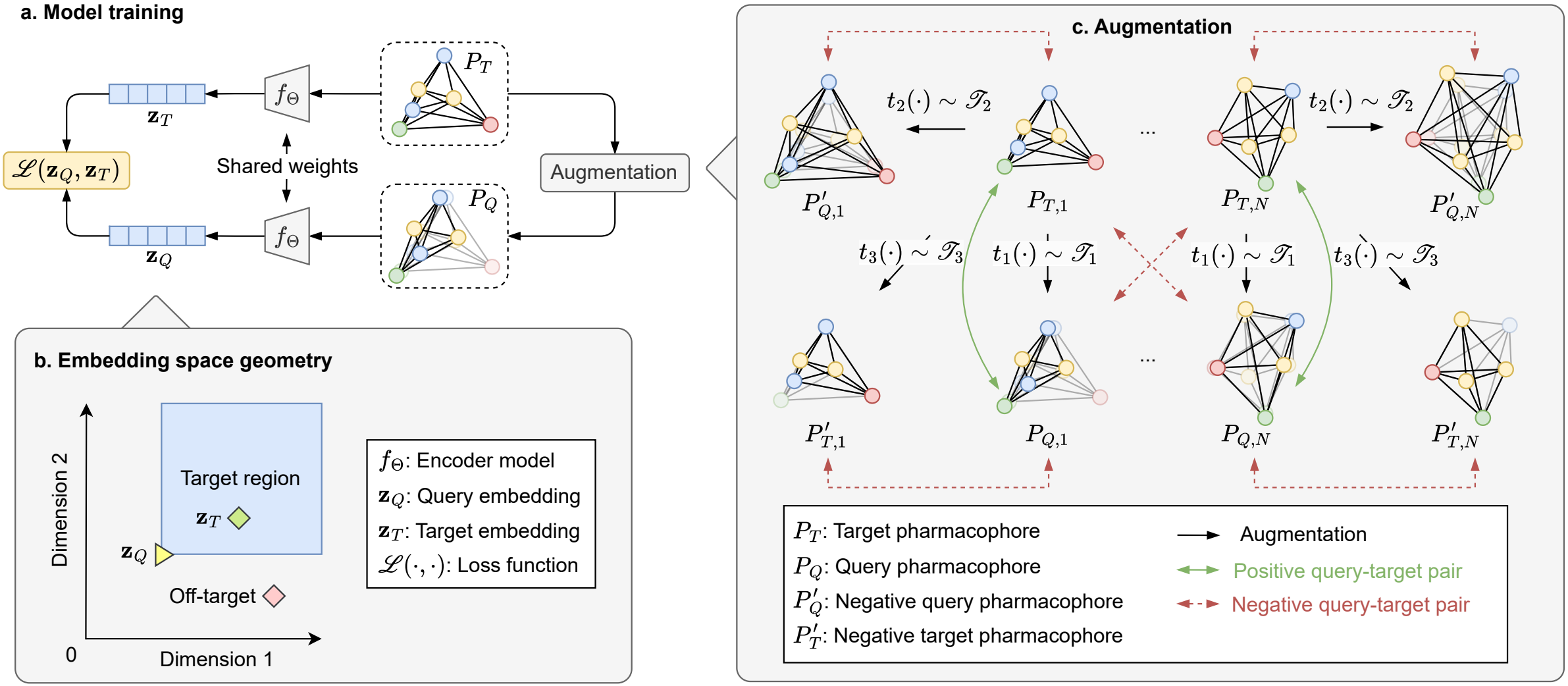
- $\forall i \in P_Q : d_i = d_{g(i)}$
- $\|\mathbf{r}_i - \mathbf{r}_{g(i)}\|_2 < r_T$
- r_T tolerance sphere radius

$$G_Q = G(P_Q),\ G_H = G(P_H),\ G_T = G(P_T)$$
$$\lambda_Q((u, v)) \approx \lambda_H((f(u), f(v)))$$

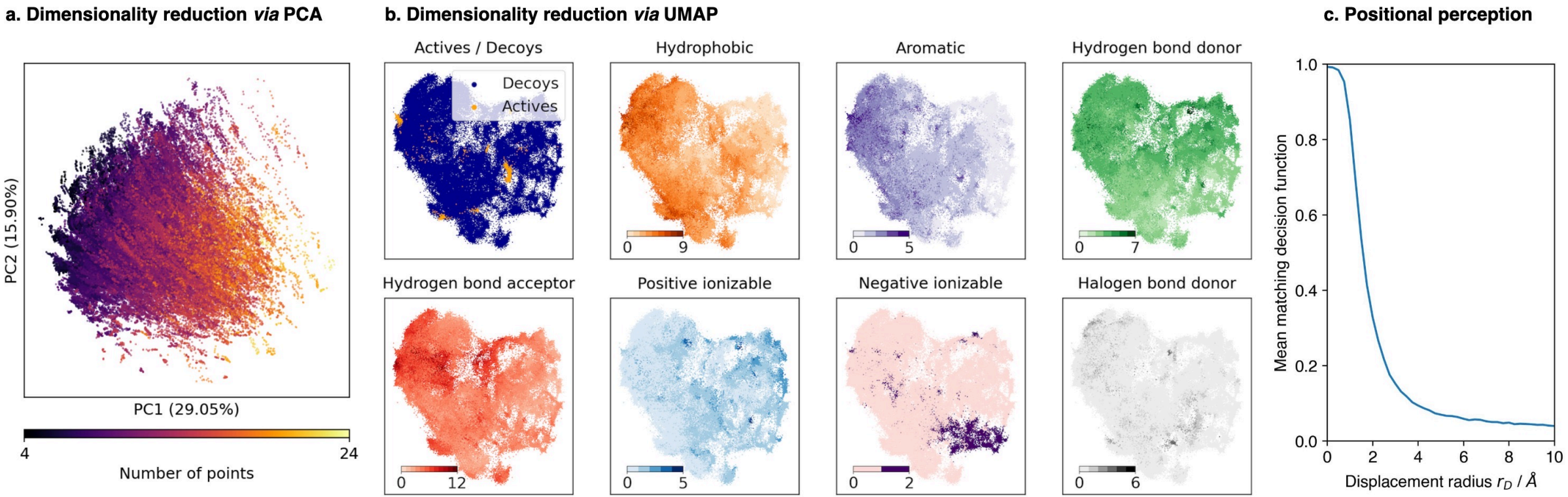
Order embedding loss encodes query target relationship:

$$\mathcal{L}(\mathbf{z}_Q, \mathbf{z}_T) = \sum_{(\mathbf{z}_Q, \mathbf{z}_T) \in Pos} E(\mathbf{z}_Q, \mathbf{z}_T) + \sum_{(\mathbf{z}_Q, \mathbf{z}_T) \in Neg} \max\{0, \alpha - E(\mathbf{z}_Q, \mathbf{z}_T)\}$$

Self-Supervised Model Training

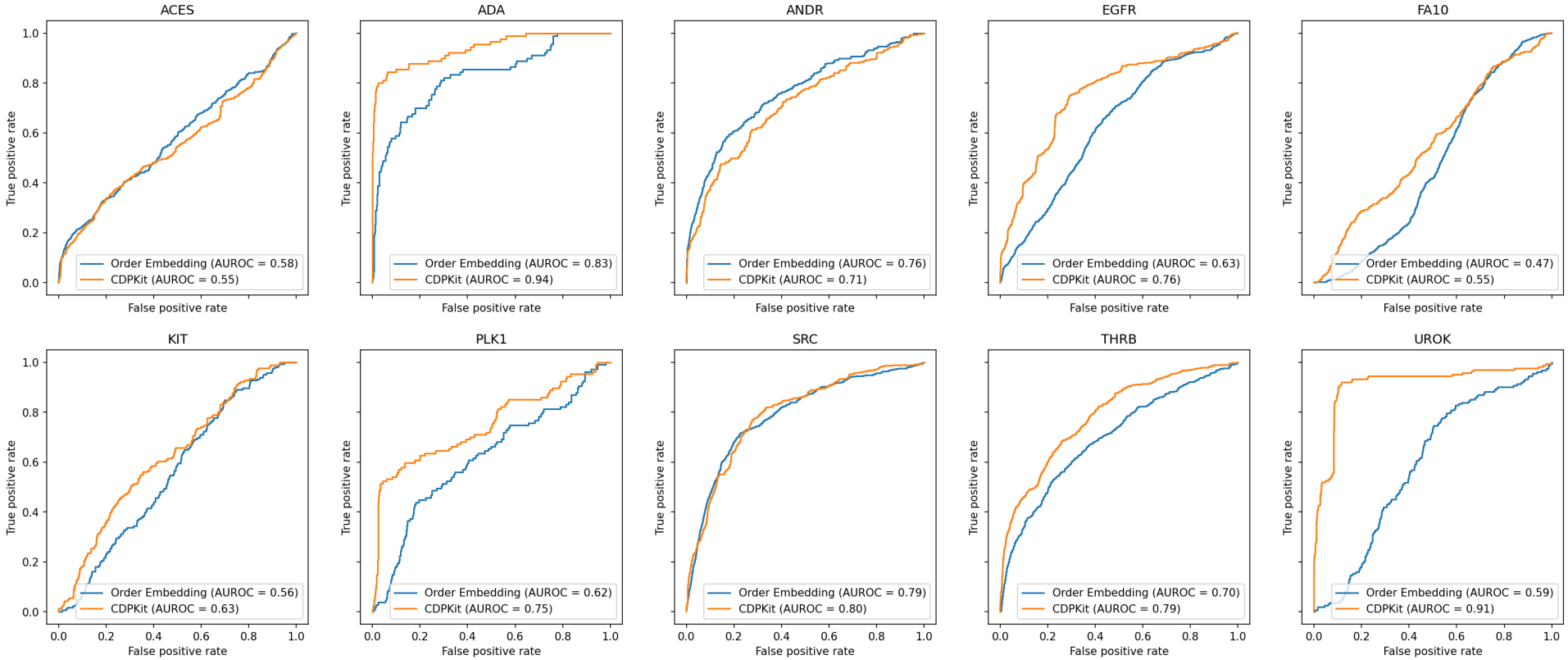


Results – Embedding Visualization



- Embedding space is structured according to graph size
- Pharmacophore feature types are represented
- 3D positional information is included

Results – Screening



	DEKOIS2.0					*LIT-PCBA*					Runtime
	AUROC	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$	AUROC	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$	per ligand (s)
PharmacoNet	**62.5**	12.3	4.4	4.2	2.9	-	-	-	3.1	-	$5.2 \cdot 10^{-3}$
PharmacoMatch (ours)	60.9	**15.1**	**5.5**	**4.9**	**3.2**	57.4	5.0	6.0	**3.5**	2.2	$\mathbf{3.3 \cdot 10^{-6}}$

Runtimes per Pharmacophore:

- Embedding time
  $67 \pm 7\ \mu s$
- Alignment time
  $66 \pm 6\ \mu s$
- Matching time:
  $0.3 \pm 0.09\ \mu s$

Summary

- Scaling pharmacophore screening to large libraries is challenging
- Pharmacophore screening can be formulated as a contrastive representation learning problem
- The learned representations can be used for efficient pre-screening

https://github.com/molinfo-vienna/PharmacoMatch

https://openreview.net/forum?id=27Qk18IZum

daniel.rose@univie.ac.at
oliver.wieder@univie.ac.at
thomas.seidel@univie.ac.at
thierry.langer@univie.ac.at

Acknowledgement
Financial support received for the Christian Doppler Laboratory for Molecular Informatics in the Biosciences by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, Boehringer-Ingelheim RCV GmbH & Co KG and BASF SE is gratefully acknowledged.