Machine Learning in Chemical Research

Katja Hansen, 12.7.2011
TU Berlin - Maschinelles Lernen

Outline

- Drug Discovery Process
- Challenges for Machine Learning in Drug Discovery
  - Chemical Descriptors
  - Confidence Estimation
  - Model Interpretation
- Further Applications
- Summary

Drug Discovery Process

1.Target Identification

Analyze the disease and the molecular mechanisms involved in order to find possible targets for drug intervention

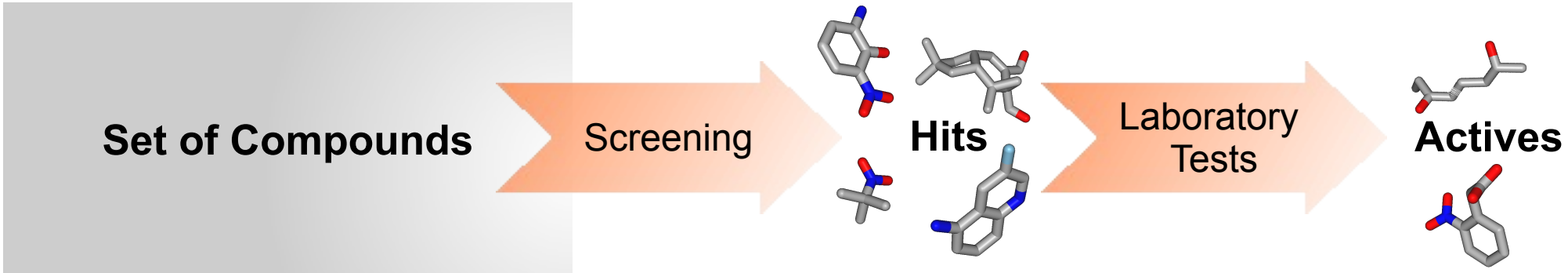
Goal:

Find a receptor that can be blocked or stimulated to alleviate a disease

2.Identify Hits: (Virtual) Screening

Goal:

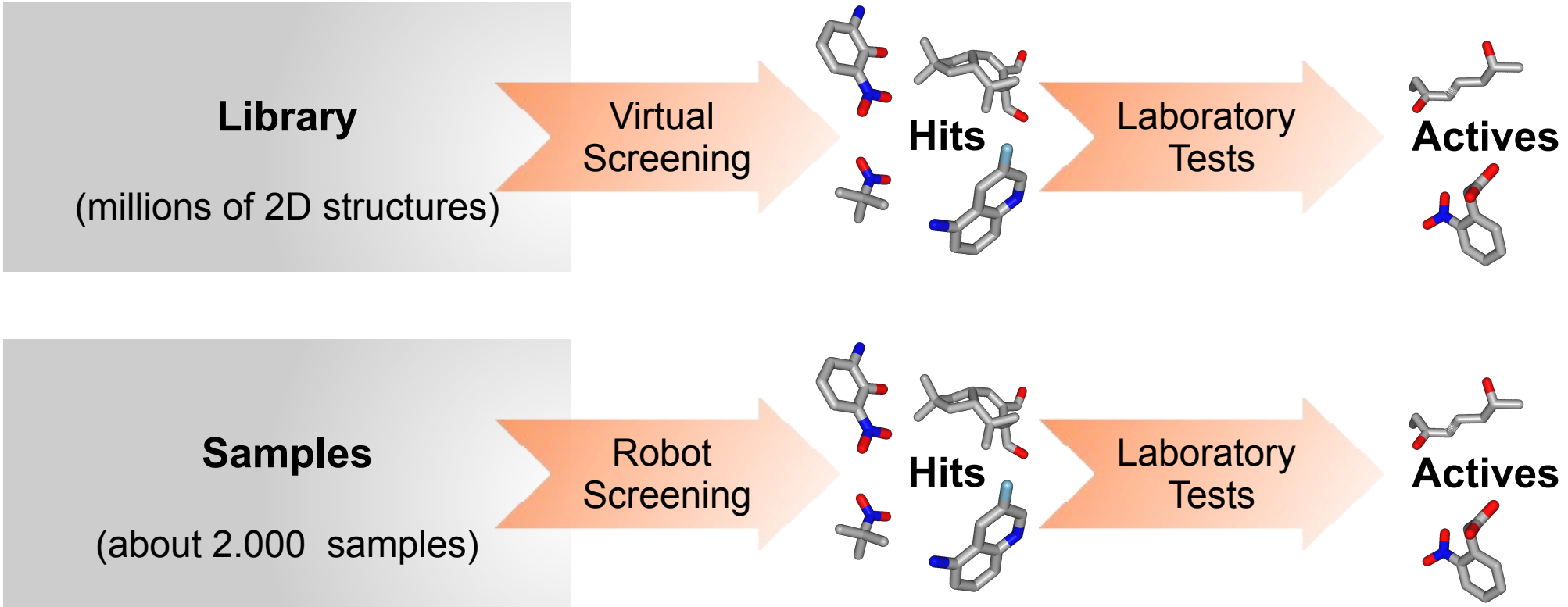
Find new compounds that bind to the identified receptor



2.Identify Hits: (Virtual) Screening

Goal:

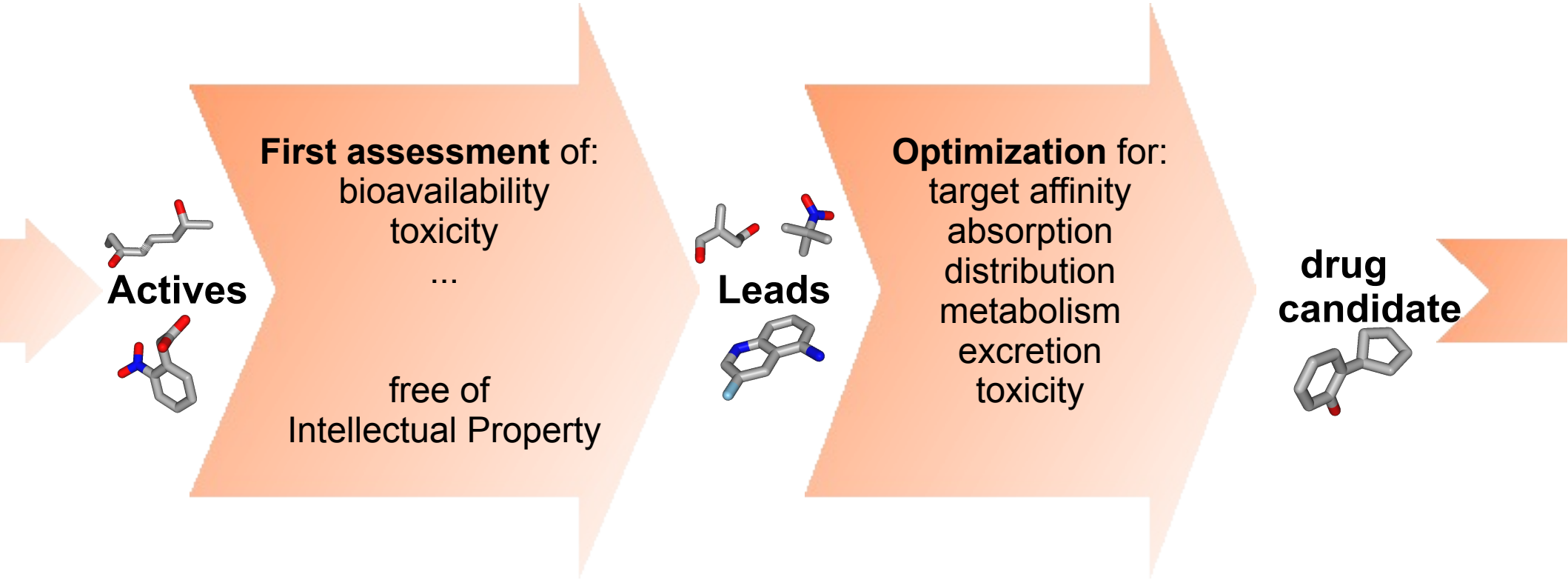
Find new compounds that bind to the identified receptor



3. From Actives to Candidate

Goal:

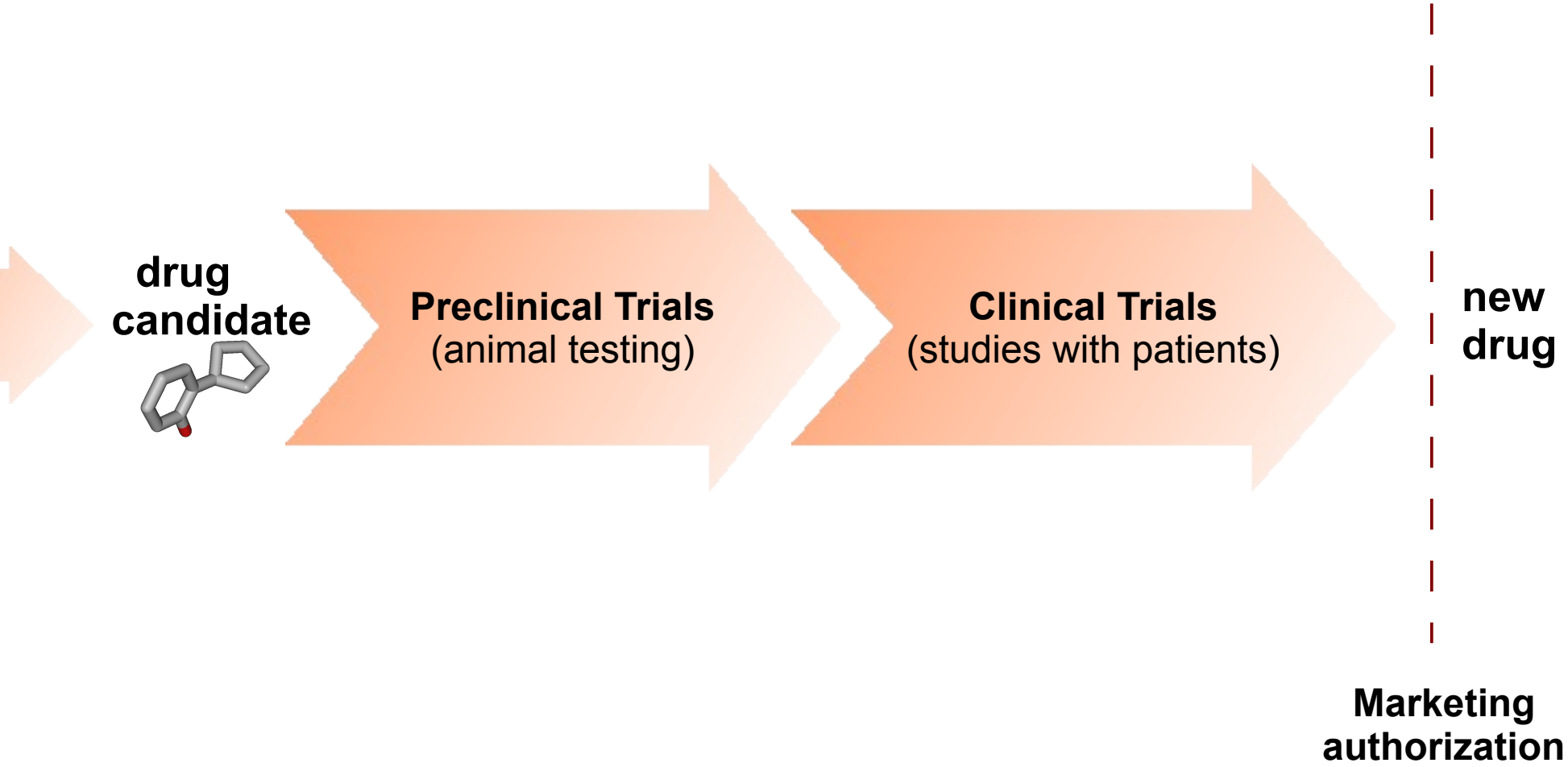
Optimize few compounds such that they have drug properties



4. From Candidate to Drug

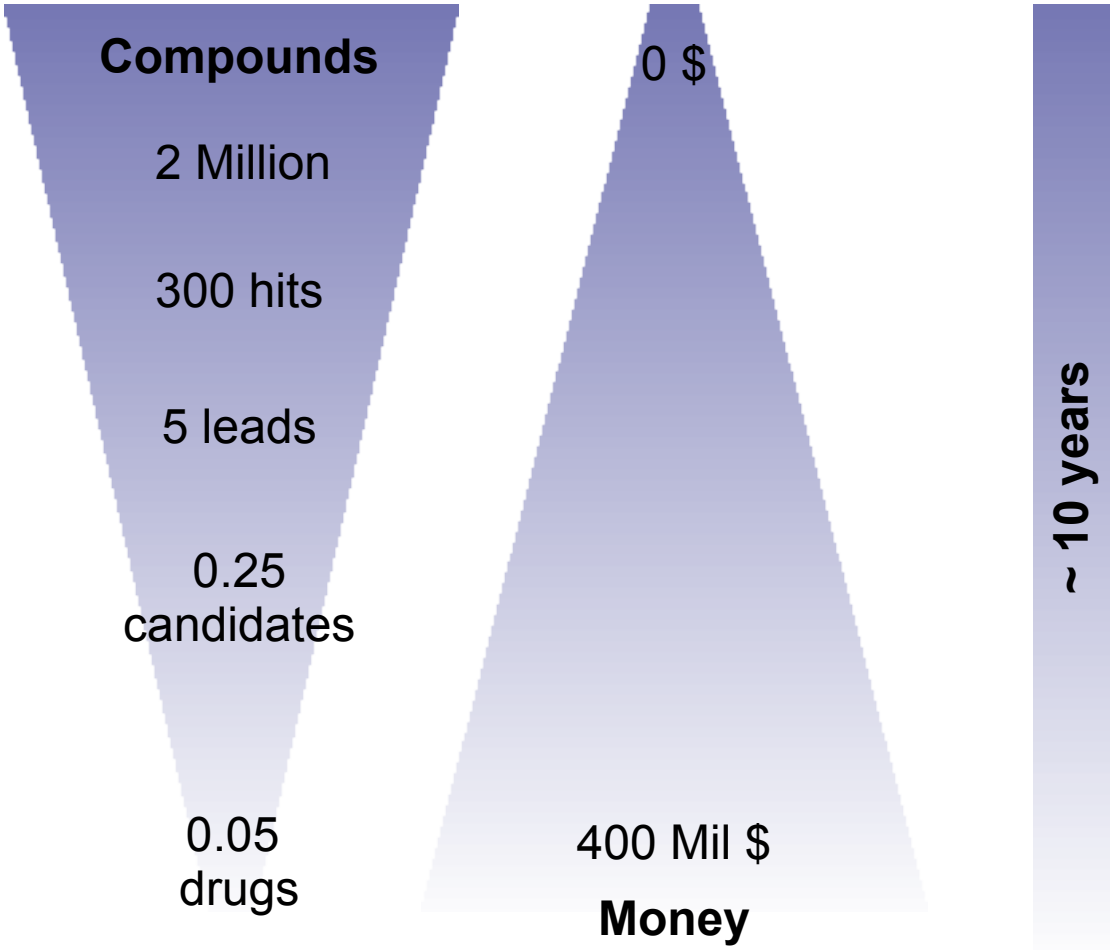
Goal:

Exclude dangerous side effects and evaluate dose



Summary Drug Discovery

- Target Identification
- (Virtual) Screening
- First Assessment
- Lead Optimization
- Preclinical trials
- Clinical Trials



Summary Drug Discovery

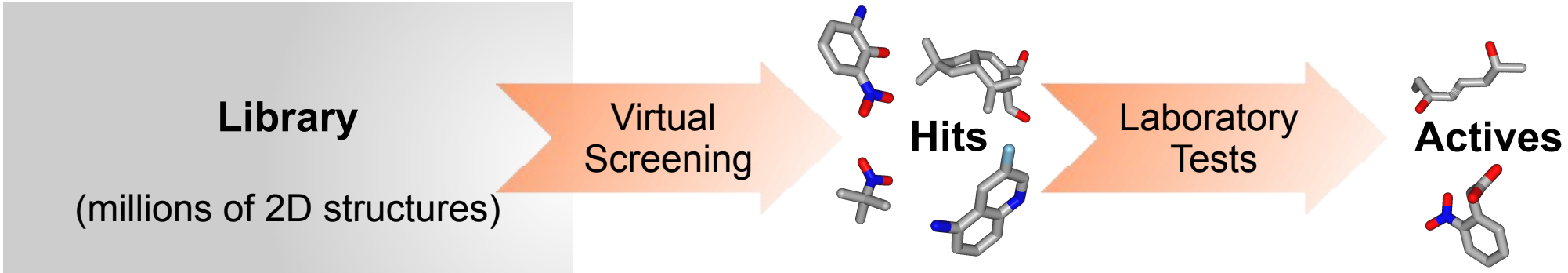
- Target Identification
- (Virtual) Screening
- First Assessment
- Lead Optimization
- Preclinical trials
- Clinical Trials

Machine Learning ?

2.Identify Hits: (Virtual) Screening

Goal:

Find new compounds that bind to the identified receptor



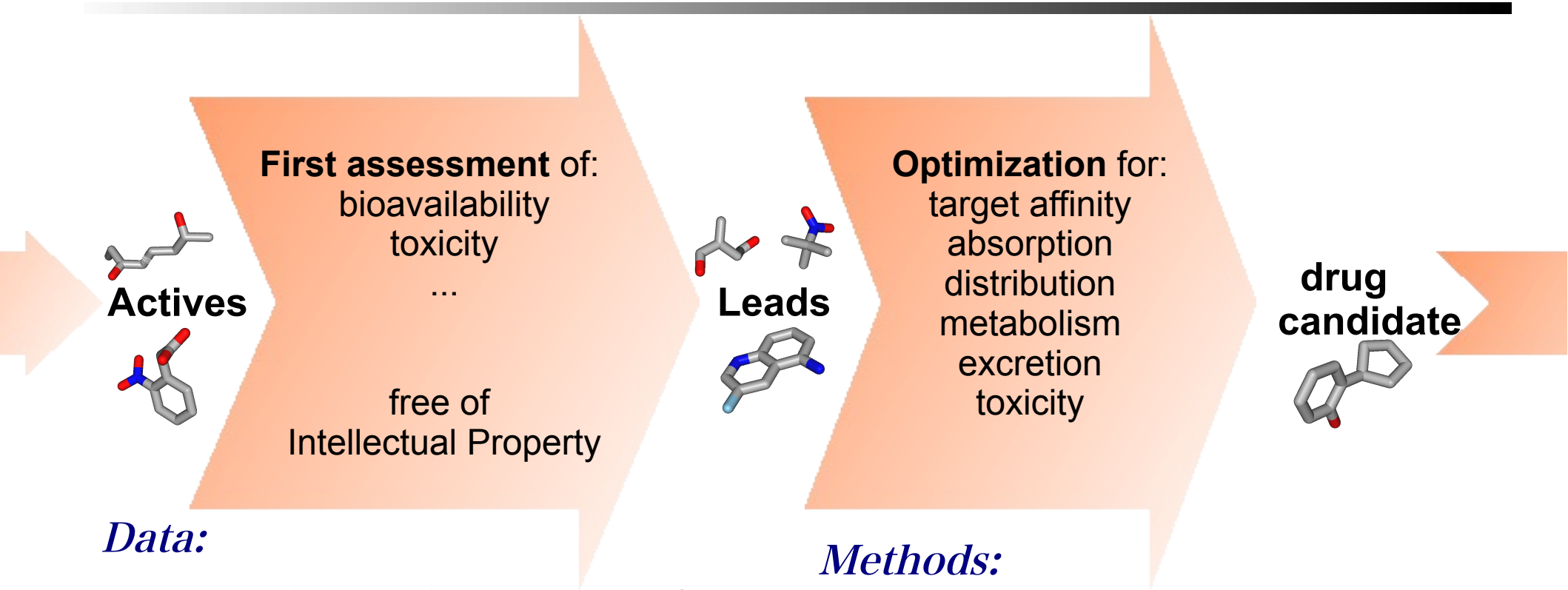
Data:

- Small number of binding drugs known from literature
- Large libraries of unlabeled chemical structures
- No negative examples

Methods:

- Semi-supervised learning
- Large-scale algorithms
- One-class detection

3. From Actives to Candidate



Data:

- Many chemical properties of various complexity
- Size and quality of data vary according to property and time
- Hardly any data available in chemical space of interest

Methods:

- Multivariate optimization
- Active learning
- Covariate shift problems (domain of applicability)
- Confidence estimation

Challenges for ML

Challenges in Drug Discovery

- Find adequate description of chemical compounds

  Descriptors

- Determine the reliability of individual predictions

  Confidence Estimation

- Understand and interpret the model

  Model Interpretation

1. Chemical Descriptors

- Need to capture characteristics of molecules for computational analysis
- Produce a vector of molecular descriptors

constitutional descriptors, counts

Topology descriptors, graph invariants

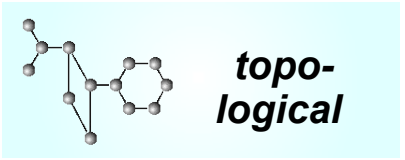
Molecular Descriptors

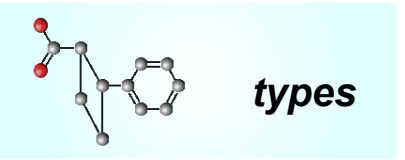
structural fragments, fingerprints

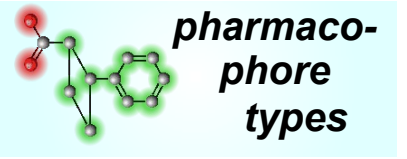
quantum-chemical descriptors, surface and volume descriptors

1. Chemical Descriptors

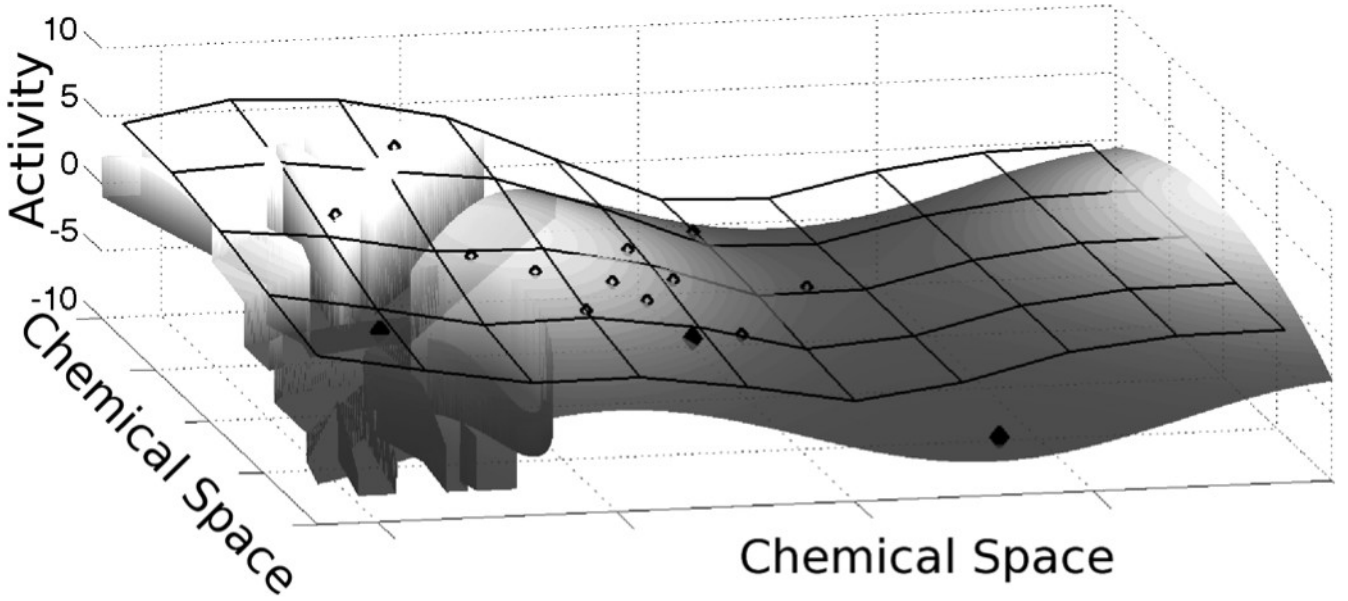
- Need to capture characteristics of molecules for computational analysis
- Use simplified molecular graph and graph kernels







2. Confidence Estimation



- The model may fail due to missing data or “activity cliffs”
- Need to know where the model gives reliable predictions

  (domain of applicability or a confidence estimate)

2. Confidence Estimation

Goal:

Estimate the prediction error of a single prediction

3 approaches on confidence estimation:

- Distance based
- Model variance
- Relevant compounds

2. Confidence Estimation

Distance Based

Idea:

- Identify neighboring compounds
- Use prediction error on neighborhood to estimate desired prediction error

2. Confidence Estimation

Distance Based

Challenges:

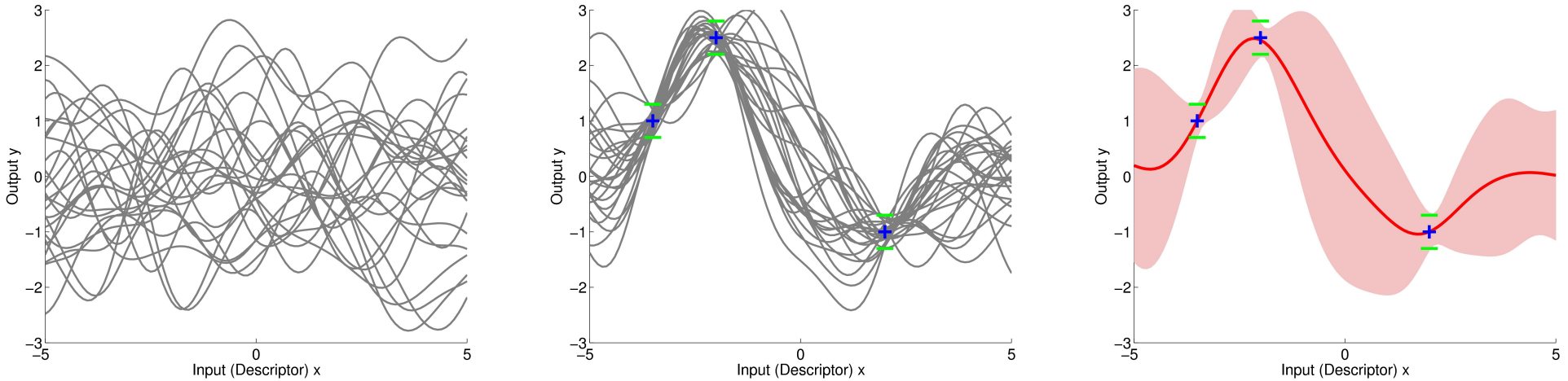
- Optimal distance metric to determine relevant neighbors:
  - sparse high dimensional space (curse of dimensionality)
  - input space or feature space
- Method of estimation
  (e.g. incorporate variance, weight distances)
- Success depends on training data set

2. Confidence Estimation

Model Variance

- For Bayesian Methods, e.g., Gaussian Process:

  Prediction is a probability distribution



- Idea:

  use predicted variance for confidence estimation

- Drawback:

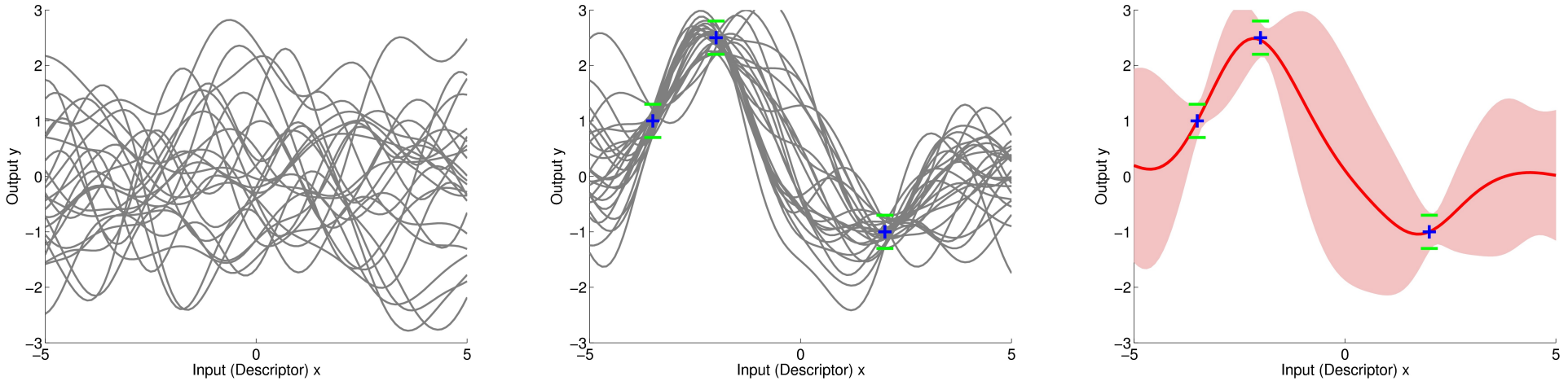
  variance and prediction error are different concepts
  the approach does not always work well in practice

2. Confidence Estimation

Model Variance

- For Bayesian Methods, e.g., Gaussian Process:

  Prediction is a probability distribution



- Idea:

  use predicted variance for confidence estimation

- Drawback:

  variance and prediction error are different concepts

  the approach does not always work well in practice

2. Confidence Estimation

Contributing Compounds

Idea:

- Explain reasoning behind single predictions to human experts
- Enable human experts to develop own confidence estimation

→ *Explain prediction by visualizing the most contributing training compounds*

2. Confidence Estimation

Calculate Contributing Compounds

Setting: Given training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Find a model to predict property y for new compound x

Identify elements of training set most relevant for prediction $\hat{y}$

2. Confidence Estimation

Calculate Contributing Compounds

Representer Theorem:

$$\hat{y} = \sum_{i=1}^{n} \alpha_i \cdot k(x, x_i) = \vec{k} \cdot \vec{\alpha}$$

each compound of the training set contributes to the prediction

Example: Gaussian Process

$$\hat{y} = \vec{k} \cdot \vec{\alpha} = \underbrace{\vec{k}\left(K + \sigma I\right)^{-1}}_{\beta} \vec{y}$$

$$= \vec{\beta} \cdot \vec{y}$$

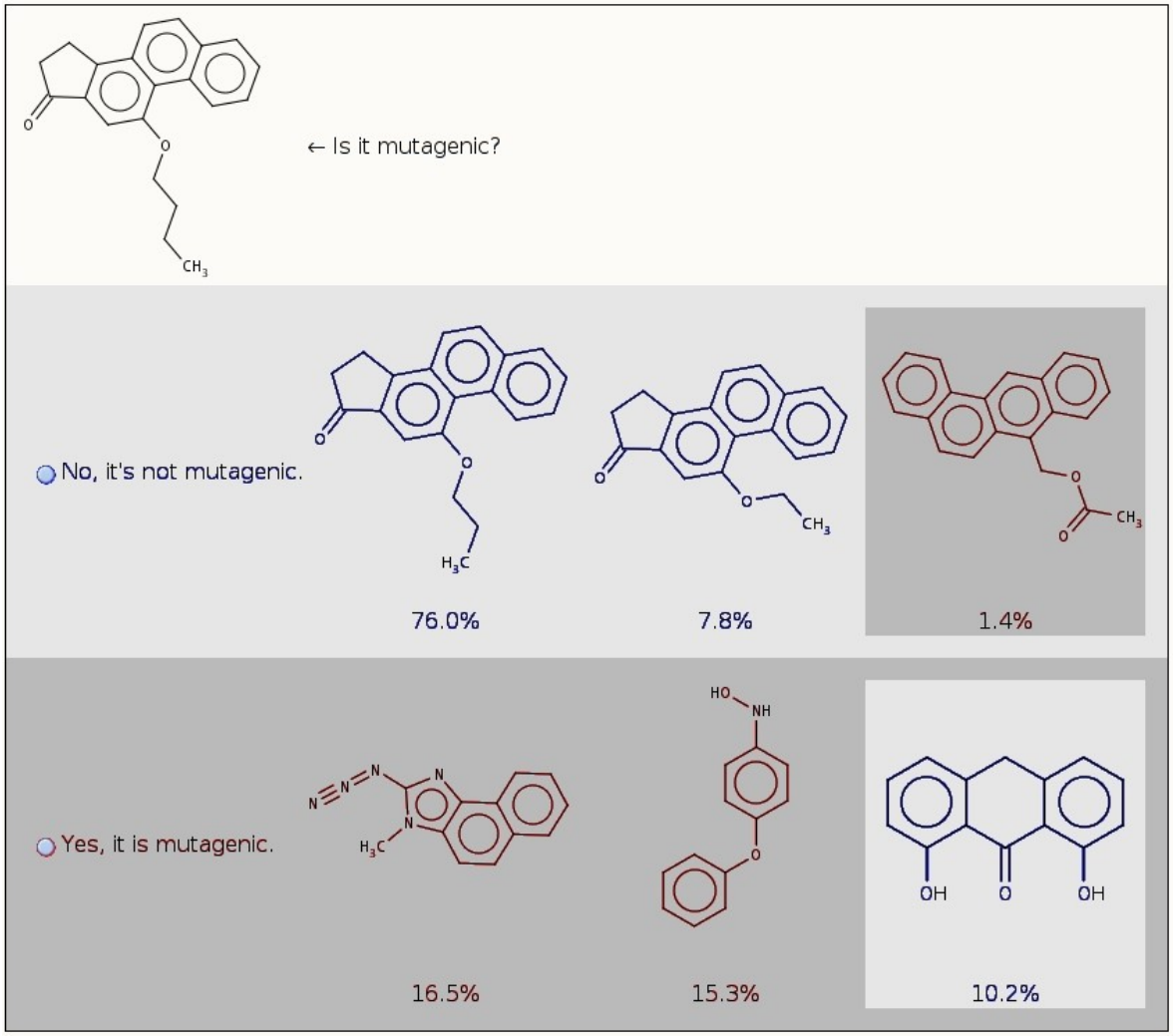
weighted sum of property values

normalized weight factor

$$\vec{\beta}_n = \frac{\vec{\beta}}{\sum_{i=1}^{n} |\beta_i|}$$

2. Confidence Estimation

Evaluate Contributing Compounds



Method:

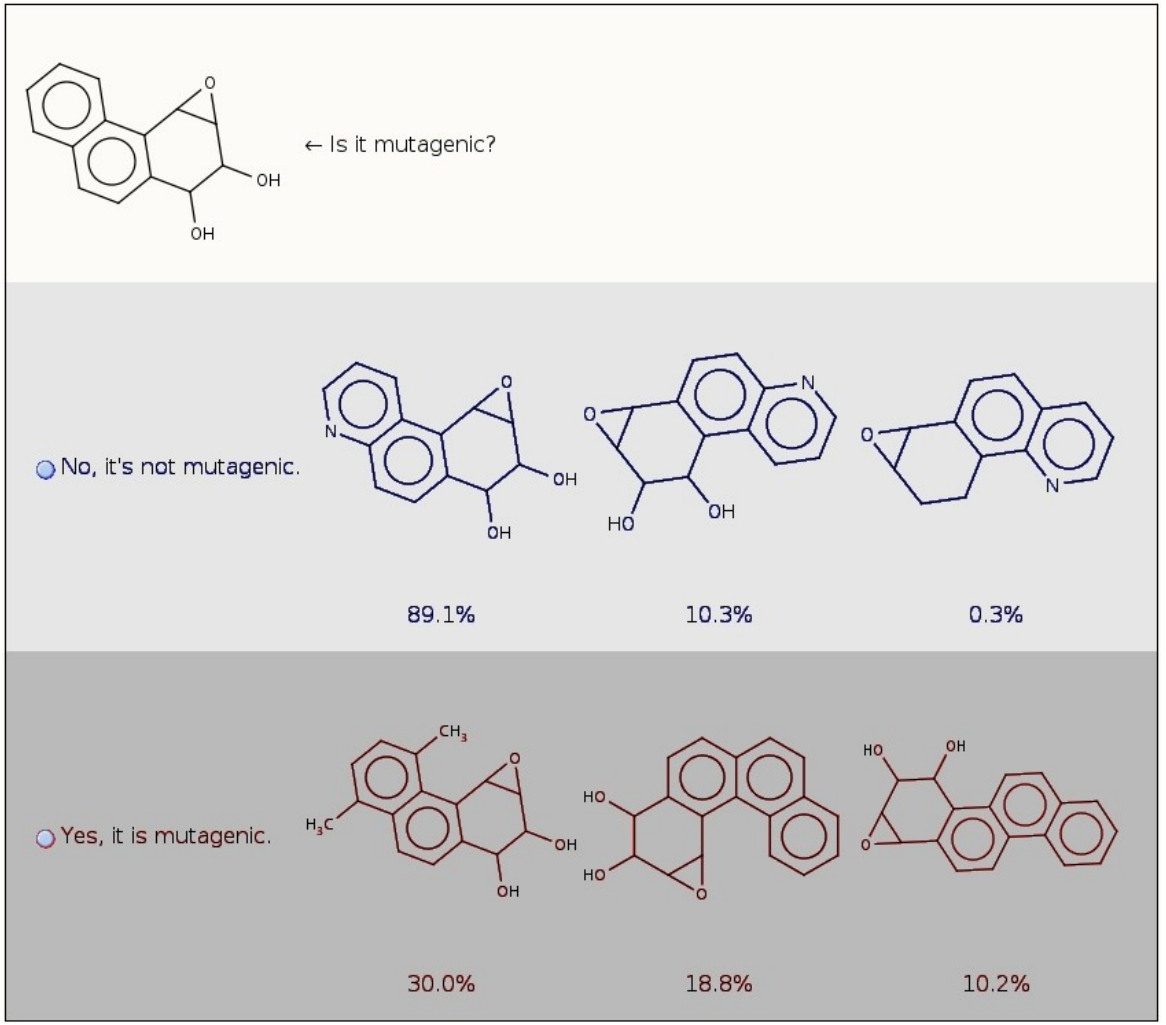
Used poll to evaluate approach:

Participants decided on reliability of mutagenicity prediction

Visualization of relevant contributing compounds or placebos

2. Confidence Estimation

Evaluate Contributing Compounds



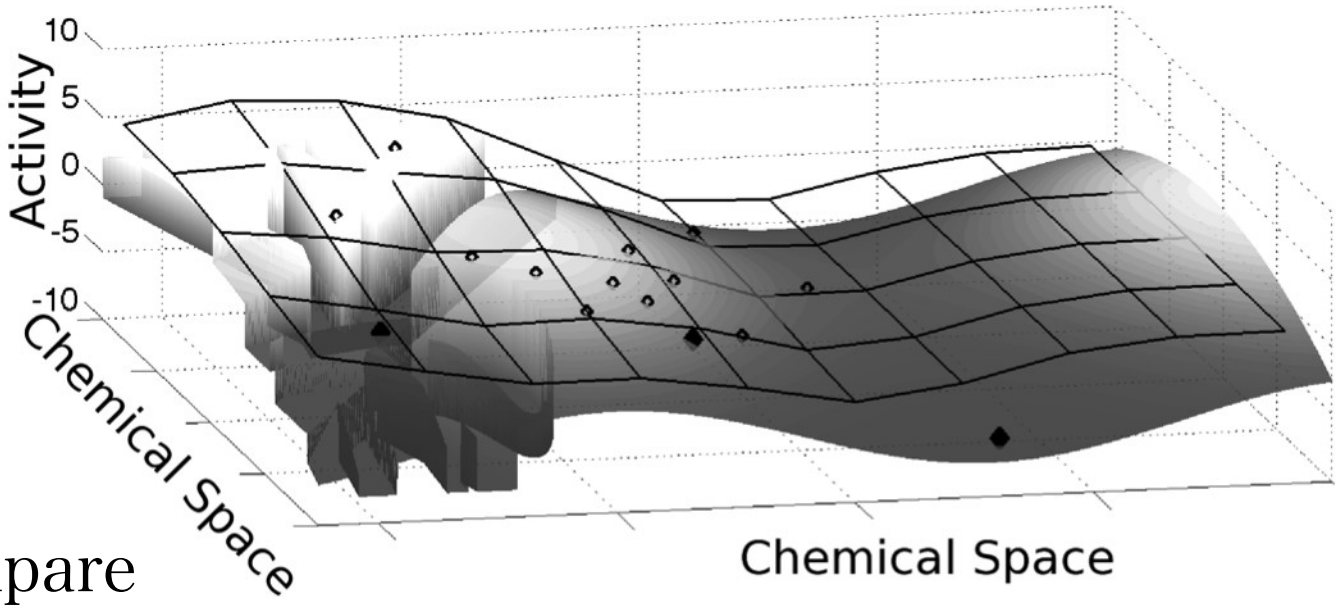
Results:

Significant improvement when using explanation compounds

2. Confidence Estimation

3 approaches on confidence estimation:

- Distance based
- Model variance based
- Contributing compounds



- No metrics to compare approaches available

3. Model Interpretation

Goal (Lead Optimization):

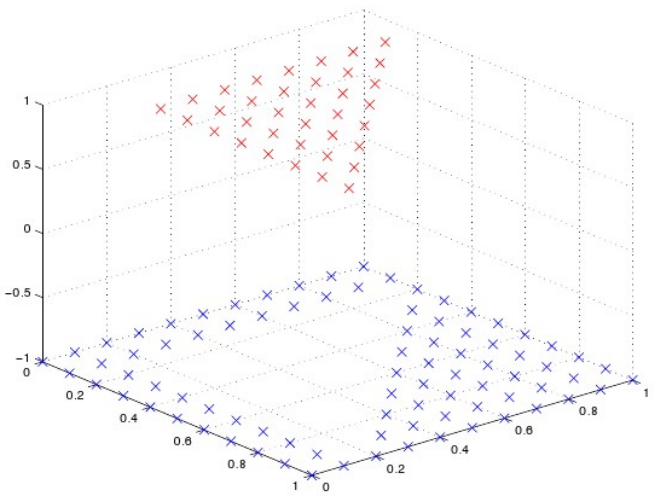
Change compound slightly to reach better property value (e.g., high binding affinity)

Idea:

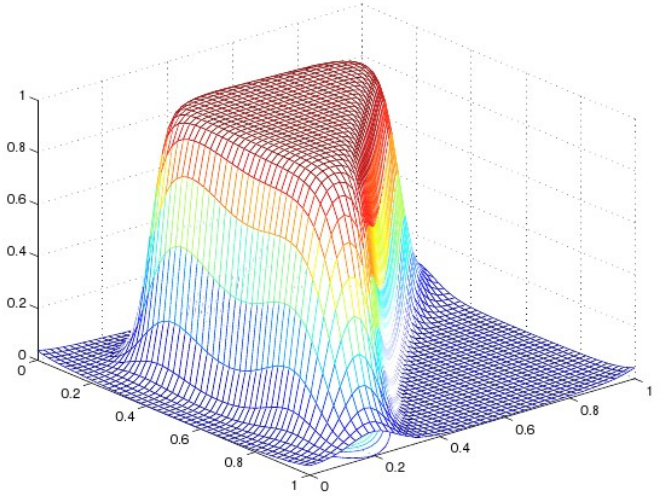
Use Machine Learning to identify relevant compound characteristics

→ *Calculate local gradients in chemical space*

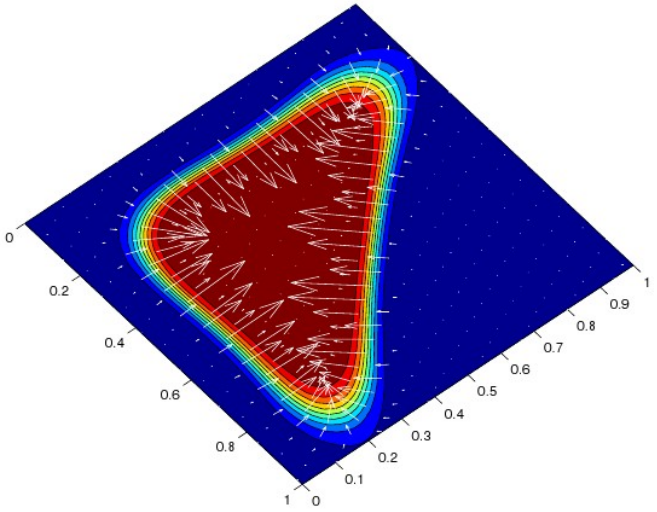
3. Model Interpretation



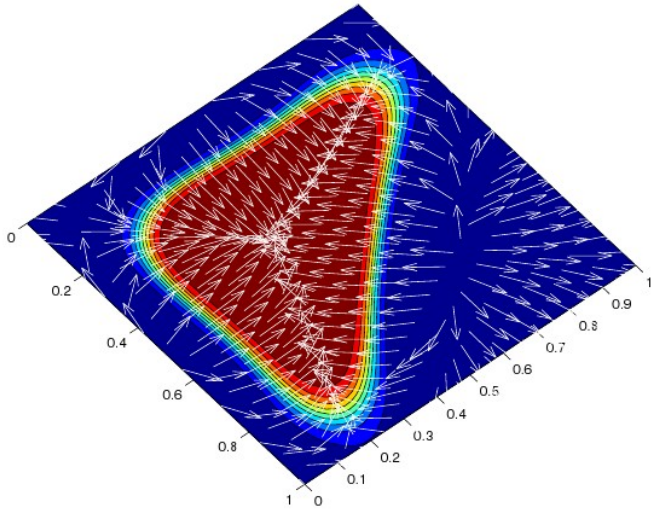
(a) Object



(b) Model



(c) Local explanation vectors

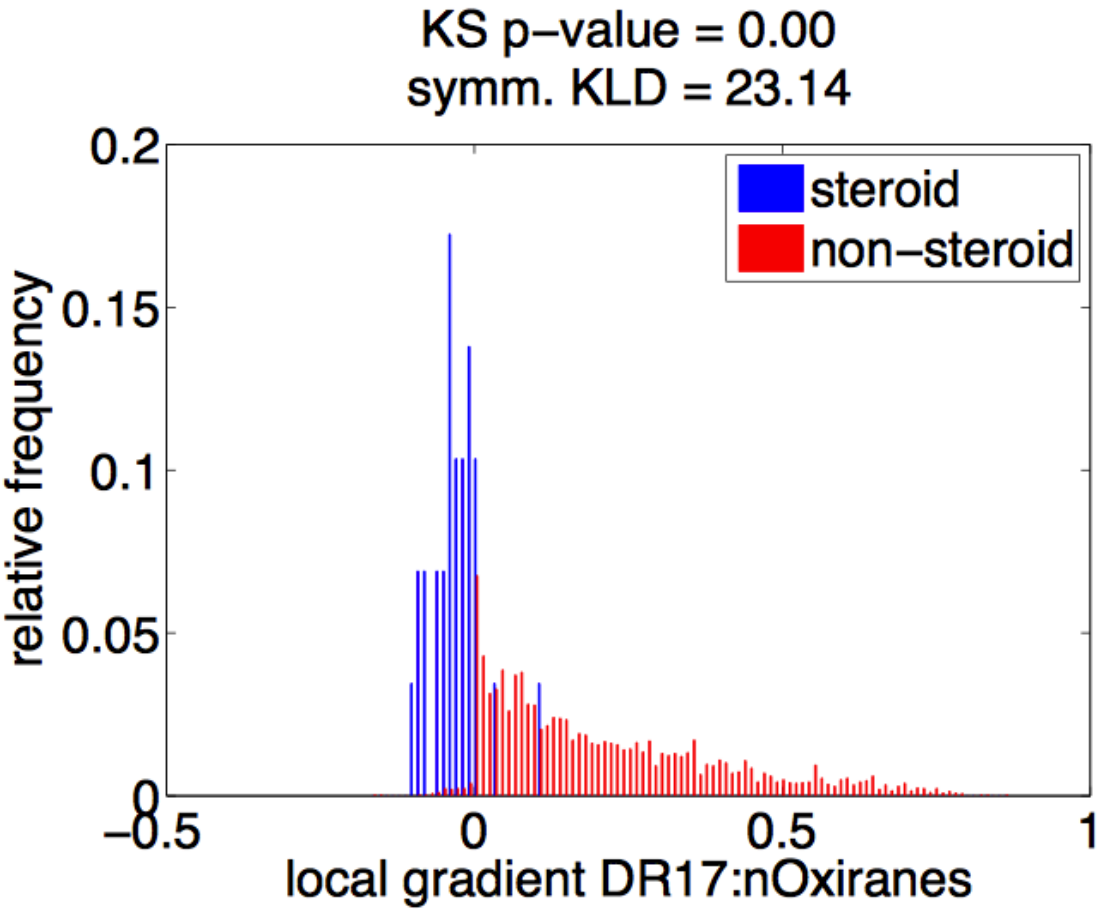


(d) Direction of explanation vectors

3. Model Interpretation

Evaluate Local Gradients

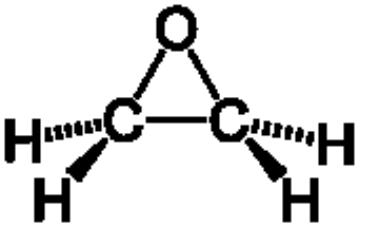
Example: Mutagenicity of Steroids



Results:

Relevant characteristics for single compound classes are detected, e.g.,

Oxiranes have no mutagenic effect on steroids



3. Model Interpretation

Calculate Locale Gradients

Example: Gaussian Process with RBF kernel

$$\hat{y}(x) = \vec{k}(K + \sigma I)^{-1}\vec{y} = \sum_{i=1}^{n} \alpha_i \cdot k(x, x_i) \qquad k(x, x_i) = e^{\frac{-\|x - x_i\|^2}{2\sigma^2}}$$

derivative of kernel function

$$\frac{\partial \hat{y}}{\partial x} = -\sum_{i=1}^{n} \alpha_i \cdot k(x, x_i) \frac{1}{\sigma^2} \|x - x_i\|$$

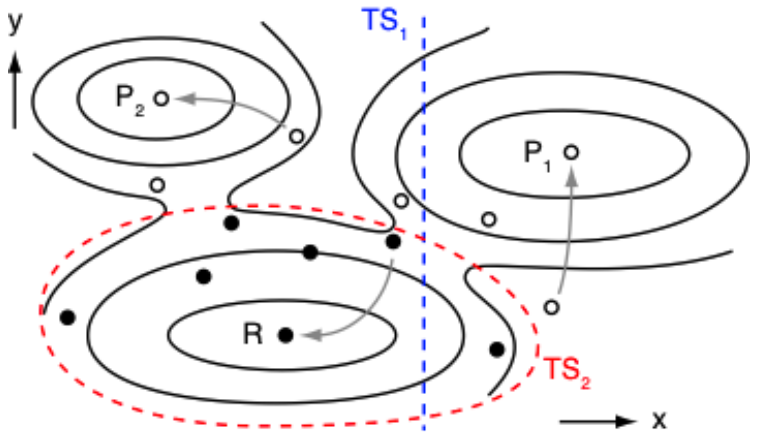
differentiate with respect to x

Further Applications

Materials Science

Goal:

- Understand the rare events leading from reactant to product using molecular dynamics
- Sample along the optimal transition state on the potential energy surface



Idea:

- Treat the setting as classification problem (product vs. reactant)
- Use local gradients of the learned machine to push sampling towards the transition state

Materials Science

Movie

Summary

- Machine Learning enhances the Drug Discovery process and saves costs, time and laboratory experiments

- Still many challenging problems, e.g.,
  - optimal compound representation
  - confidence estimation
  - model interpretation

- There is room for improvement and I am looking forward to your ideas...

References

- Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors* John Wiley & Sons, Ltd., **2000**
- Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V. & Tetko, I. V. *Applicability domain for in silico models to achieve accuracy of experimental measurements* Journal of Chemometrics, **2010**, 24, 202-208
- Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N. & Müller, K.-R. *Benchmark Data Set for in Silico Prediction of Ames Mutagenicity* J. Chem. Inf. Model., **2009**, 49, 2077-208
- Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K. & Müller, K.-R. *How to Explain Individual Classification Decisions* JMLR, **2010**, 11, 1803-1831