Exam time was 90 minutes and there were german translations which I did copy.

Karlsruhe Institute of Technology
Department of Informatics
T.T.-Prof. Dr. Pascal Friederich

04.08.2025

Machine learning for the natural sciences
Maschinelles Lernen für die Naturwissenschaften
SS 2025

Name:

KIT ID (uxxxx):

Matrikel number:

Subject of study (Studiengang):

Please do not fill:

Exercise 1	Exercise 2	Exercise 3.1	Exercise 3.2	Total	Mark
(of 13)	(of 8)	(of 19)	(of 20)	(of 60)	

Remarks

- Please check the exam for completeness (18 pages).
- Place your student ID card visibly on the desk.
- Do not use red or green color.
- Turn off your cell phone and put it in your pocket.
- The use of unauthorized aids (books, cheat sheets, cell phone, calculator, lecture slides, etc.) is considered an attempt to cheat and will result in a grade of "not sufficient" (5.0) for the exam and may lead to exmatriculation.
- Please write clearly. Texts and answers that are not readable cannot be graded.
- Limit yourself to the required information and keep your answers short and precise. You may answer in bullet points if possible.
- Save time by only reading the question in one language.
- You have 15 minutes to read the exam and clarify questions of understanding. No questions may be answered after that time.

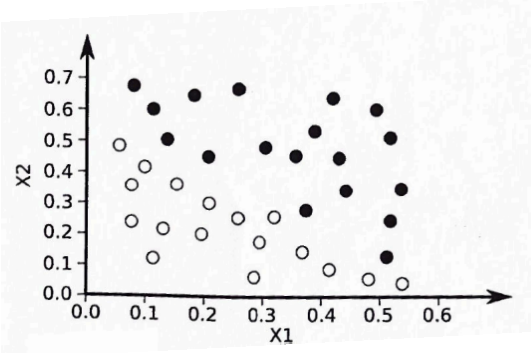
All the best! Viel Erfolg!

Exam time was 90 minutes and there were german translations which I did copy.

1. Multiple Choice (1 point per question, 13 points in total)

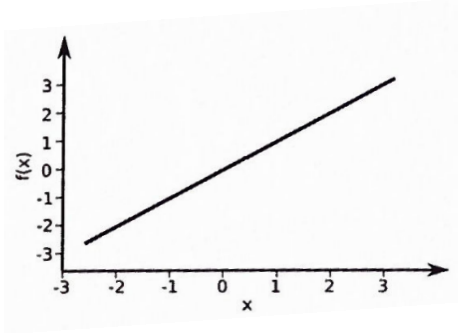
Please carefully read the questions and answers. Check all correct answers (**multiple correct answers per question are possible**). Only complete and fully correct answers give a point. Partially correct or partially wrong answers give 0 points.

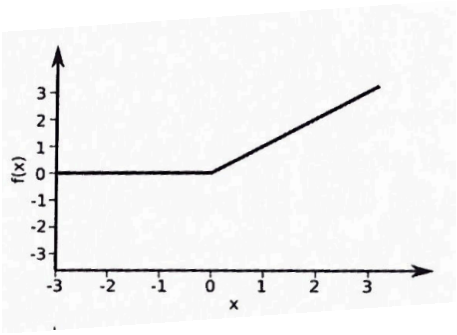
1.1. **Decision trees.** You have the following dataset with features X_1 and X_2 and the labels black and white. Which split yields the maximum impurity reduction? **(1 point)**

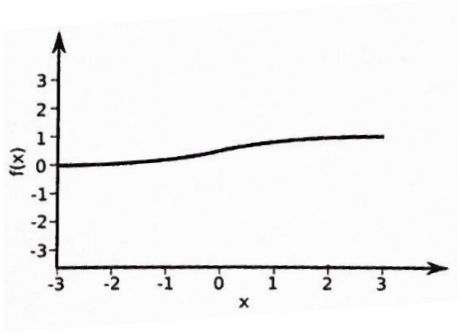


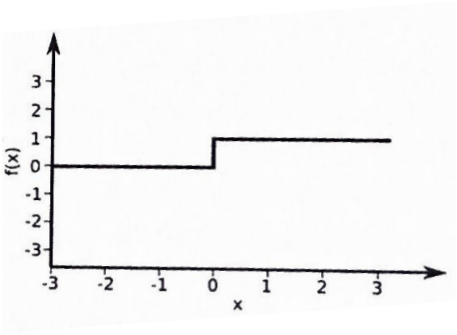
- ☐ $X_1 > 0$
- ☐ $X_2 < 0.5$
- ☐ $X_1 < 0.3$
- ☐ $X_1 + X_2 > 0.6$

1.2. **Activation functions.** Which of the functions below is conceptually most suitable for the output layer of a neural network for binary classification that should be trained with backpropagation? **(1 point)**

- ☐ 

- ☐ 

- ☐ 

- ☐ 

Exam time was 90 minutes and there were german translations which I did copy.

1.3. **Activation functions.** Which statements about the sigmoid activation function are correct? **(1 point)**

- ☐ The sigmoid activation function is piecewise linear.
- ☐ Sigmoid is primarily used for handling sequential data, such as time series or natural language processing tasks.
- ☐ The sigmoid activation function is defined as $f(x) = 1/(1 + e^{-x})$
- ☐ The sigmoid activation function is computationally efficient compared to other activation functions like ReLU.
- ☐ In the output layer of a neural network, sigmoid is commonly used for regression problems.
- ☐ At small and large inputs, no gradient is defined, which leads to the vanishing gradient problem.

1.4. **Random Forests.** How does a random forest improve upon a single decision tree model? **(1 point)**

- ☐ Random forests are deeper than individual decision trees.
- ☐ Random forests combine multiple weak models into a strong model.
- ☐ Random forests train each tree on the same subset of the data but uses different weight initializations.
- ☐ Random forests have a lower variance than individual decision trees.

1.5. **electrocardiogram (ECG) classification.** Which model or models are appropriate for classifying ECG sequences of variable lengths, e.g. to classify specific anomalies in the signal? **(1 point)**

- ☐ U-Net
- ☐ RNN
- ☐ CNN
- ☐ LSTM

1.6. **CNNs.** A convolution layer in a CNN has 4 filters, each with a size of 5×5 and a stride of 2. The input to this layer has 3 channels. How many trainable parameters does this convolution layer have (ignore bias parameters)? **(1 point)**

- ☐ 25
- ☐ 75
- ☐ 100
- ☐ 300

Exam time was 90 minutes and there were german translations which I did copy.

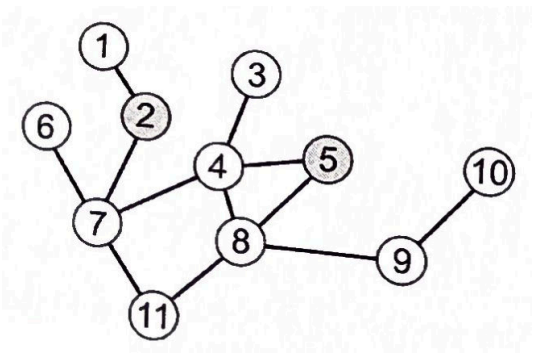
1.7. **ML basics.** Which statements are true about the bias-variance-tradeoff? **(1 point)**

- ☐ High bias and high variance typically do not occur together, i.e. bias and variance are negatively correlated.
- ☐ In order to reduce the variance of a model, its capacity should be decreased, e.g. by using regularization.
- ☐ Compared to a single decision tree, random forests have a higher variance.
- ☐ Compared to a single decision tree, random forests have a higher bias.

1.8. **Neural network potentials.** Which of the following statements are true? **(1 point)**

- ☐ Energies in neural network based potentials can be obtained by computing the derivatives of the predicted forces with respect to the atomic coordinates.
- ☐ If ground truth forces are also available during training time, they can be used as an additional term in the loss function which can lead to a higher accuracy of the neural network potential.
- ☐ Global aggregation (or read-out) function of node vectors is needed when graph neural networks are used as neural network potentials, because energy labels are only available for the full system but not for individual atoms.

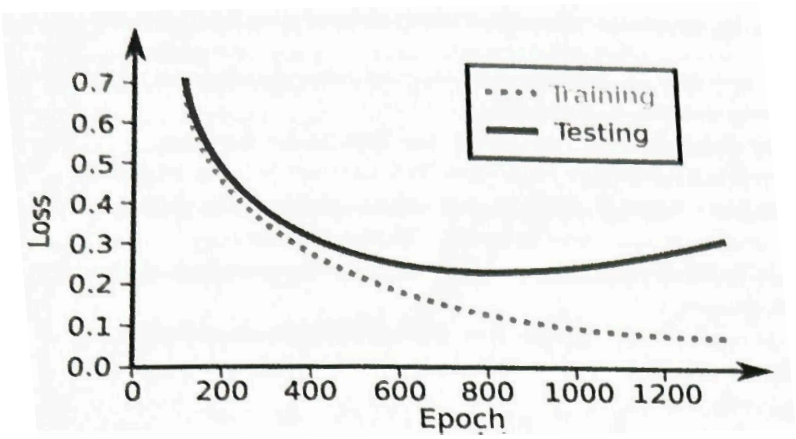
1.9. **Graph Convolutions.** You are applying a graph convolutional network (GCN) with 2 layers to the the illustrated input graph. Which of the following nodes can contain information from both of the bold nodes (node 2 and 5) after the message passing process? **(1 point)**



- ☐ 3
- ☐ 7
- ☐ 8
- ☐ 11

Exam time was 90 minutes and there were german translations which I did copy.

1.10. **Neural networks.** Suppose your colleagues are trying to predict the elasticity of given materials using a graph neural network with a training testing split of 95 : 5. They show you the result plotted in the graph below. What do you recommend your colleagues to do? **(1 point)**



- ☐ You should train the model with more epochs to improve the test loss.
- ☐ The model is suffering from overfitting and needs more regularization.
- ☐ The model should not be used because it did not learn anything at all.
- ☐ Using a training: testing split of 50 : 50 would probably reduce the test loss substantially.
- ☐ Generating more training data would probably close the gap between training and testing and thus improve the model.
- ☐ A more complex model would likely reduce the test loss.

1.11. **Neural networks.** You train a neural network on a given dataset and you are unhappy with the performance. You do a learning curve analysis, i.e. you analyse the training error and validation error of the neural network as a function of the training set size. You see that after an initial decrease in error, both errors remains stable and more or less identical above an intermediate training set size. How could you improve the model's performance? **(1 point)**

- ☐ Generate more data so you can increase the training set size.
- ☐ There is nothing you can do, as you reached the intrinsic noise level of the data.
- ☐ Adjust the hyperparameters to have a model with higher capacity to reduce the bias of the model
- ☐ Increase the resolution and/or completeness of the input representation.

Exam time was 90 minutes and there were german translations which I did copy.

1.12. **Bayesian optimization.** Which of the following statements is/are correct about Bayesian optimization (BO)? **(1 point)**

- ☐ BO is a suitable algorithm for problems where the objective function evaluation is expensive.
- ☐ BO is a local optimization method, similar to gradient descent. Momentum can be used to overcome local barriers.
- ☐ The objective function does not have to be differentiable in order for BO to be usable.
- ☐ The computational cost of each BO step scales cubicly with the number of training data points.
- ☐ BO can be parallelized by evaluating the objective function multiple times in parallel. However, the overall efficiency of the algorithm will be reduced.

1.13. **CNNs.** You are training a CNN model for semantic segmentation of biological microscopy images, i.e. you have images and you want to predict binary labels for each pixel which say whether the pixel belongs to a cell or not. Which of the following statements is true? **(1 point)**

- ☐ A pre-trained and fixed ResNet model can be used to extract vector representations of the input images which help to predict image labels.
- ☐ A U-Net architecture is not useful here because the resolution in the bottleneck layer is too low to reconstruct an image with full input resolution.
- ☐ A U-Net architecture can be used here because the input and output have the same shape (resolution).
- ☐ Data augmentation cannot be used here, as the labels (segmentation masks) change under rotation and scaling of the input images.
- ☐ A recurrent neural network has to be used, as the output size (the number of pixels that is part of the cell) changes from image to image.

Exam time was 90 minutes and there were german translations which I did copy.

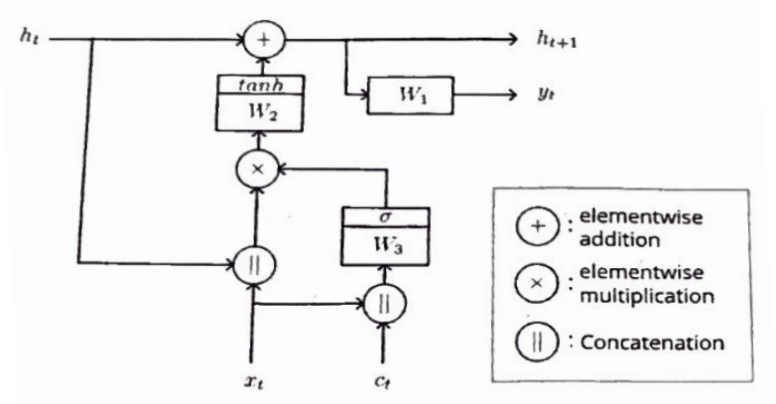
2. Derivations (8 points in total)

2.1. **Bayesian optimization.** You are maximising a function using Bayesian optimization. You have a few initial observations and fitted a Gaussian process model which results in a mean prediction $\mu(x)$ and an uncertainty interval $\sigma(x)$. Explain the following acquisition functions in terms of exploration and exploitation, and evaluate whether they are good choices or not. Explain your judgements. Assume that the maximum of the acquisition function is chosen for the next experiment. $u_1 = \mu(x), u_2 = \sigma(x), u_3 = \sigma(x) - \mu(x), u_4 = \mu(x) + \sigma(x)$. **(2 points)**

2.2. **Quantification of purity gain.** In decision trees, impurity of a sample set X can be quantified using the entropy function: $I(X) = -\sum_{c \in \{\text{red, blue}\}} p_c \log_2(p_c)$, where p_c is the proportion of class c in the set X. Consider a dataset X with $r_1 = 4$ red balls and $b_1 = 8$ blue balls. A split divides the data into two subsets: (1) $X_1 : r_{21} = 4$ red, $b_{21} = 0$ blue; and (2) $X_2 : r_{22} = 0$ red, $b_{22} = 8$ blue. Determine the reduction of impurity from this split. Express the solution in $\log_2(i)$ with integer numbers i. **(3 points)**

Exam time was 90 minutes and there were german translations which I did copy.

2.3. **RNN Architecture.** The given figure illustrates a novel recurrent neural network layer architecture. Each layer receives three input vectors: the input vector $x_t \in \mathbb{R}^7$, the condition vector $c_t \in \mathbb{R}^5$ and the previous hidden vector $h_t \in \mathbb{R}^{20}$. Each layer results in 2 output vectors: The output vector $y_t \in \mathbb{R}^9$ and the updated hidden vector h_{t+1}. The layer makes use of three learnable weight matrices W_1, W_2, W_3. Determine the shapes of the weight matrices W_1, W_2, W_3 and give them as a tuple (n, m). **Hint:** Assume all signals are column vectors and that input vectors are multiplied with the weight matrices from the right-hand side: $b_i = \sum_j W_{ij} \cdot a_j$. Please ignore bias parameters. **(3 points)**



Exam time was 90 minutes and there were german translations which I did copy.

3. Open questions

3.1. Machine learning basics (19 points in total)

3.1.1. Decision Trees and Ensemble Methods.

a) List two hyperparameters in a decision tree and explain their impact on the model's capacity.
b) Consider a decision tree that overfits the training data, as indicated by very high training accuracy but low test accuracy. Describe the process of pruning and how it can improve the model's performance on unseen data.
c) Describe the concept of bagging in ensemble methods. Discuss how this method addresses bias and variance and which requirements have to be fulfilled for that.
(3 points)

Exam time was 90 minutes and there were german translations which I did copy.

3.1.2. **Bayesian optimization.**
a) What is the basic algorithm of Bayesian optimization (no formal algorithm required, just a description in form of keywords)?
b) What is Bayesian optimization frequently used for: List one application in machine learning and one application in materials science. For each of these two applications, list the optimization parameters and the objective function. **(3 points)**

3.1.3. **Neural networks.** What happens to the depth (2-10), der size of the hidden layers (10-100), the number of training epochs (50-1000), and the L1 regularization parameter (0-1) of a neural network if these four hyperparameters are determined based on a minimization of the **training** loss? **(2 points)**

Exam time was 90 minutes and there were german translations which I did copy.

3.1.4. **Transfer learning.**

a) Please explain the idea of transfer learning and give an application example.
b) How would you apply transfer learning to LSTMs (for text classification) and graph neural networks (for molecular property prediction)? **(4 points)**

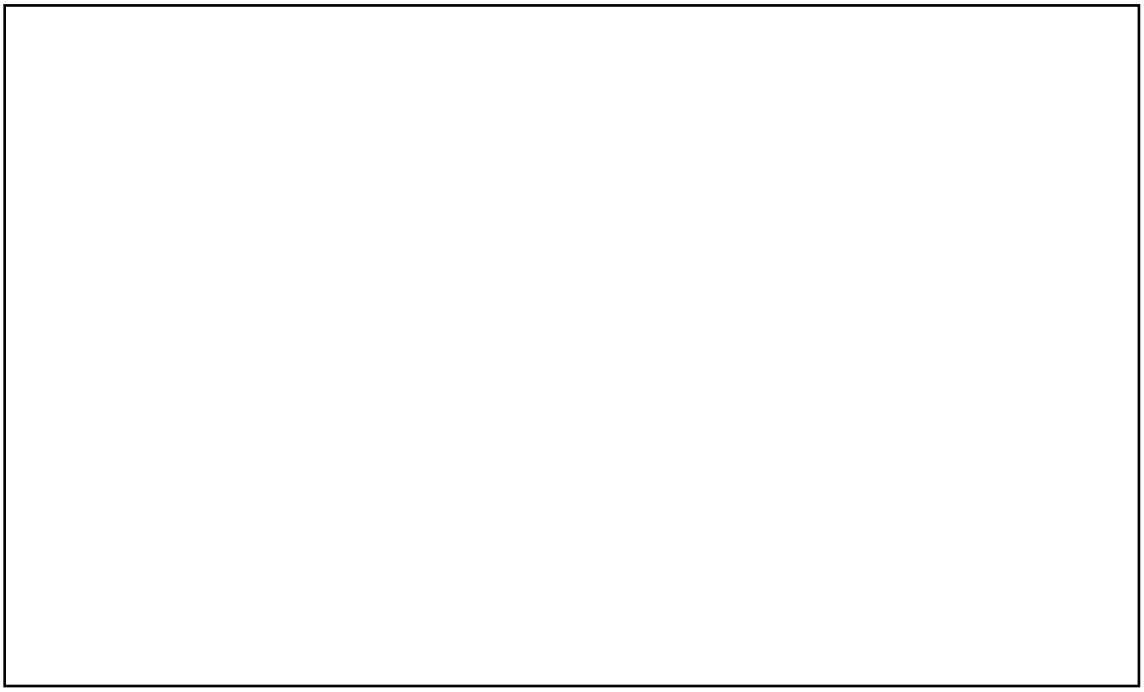
Exam time was 90 minutes and there were german translations which I did copy.

3.1.5. **VAEs vs. AEs.**

a) Briefly explain how variational autoencoders (VAE) differ from standard autoencoders (AE) regarding the loss function, the structure of the latent space, and possible applications.

b) What is the purpose of the KL divergence term in the loss function, and how does it affect the latent space?

c) Imagine we have a VAE trained on the MNIST dataset (images of handwritten digits from 0 to 9). Given two images, A and B, we can linearly interpolate their latent representations L_a and L_b generating the set of points $S : \{L_i \mid L_i = x_i * L_a + (1 - x_i) * L_b, x_i \in [0, 1]\}$. If we reconstruct S using the VAE's Decoder, we will see a smooth transition from the digit in image A to the digit in image B. Describe and explain what issues might arise if we used a standard AE instead. **(3 points)**



Exam time was 90 minutes and there were german translations which I did copy.

3.1.6. **Active learning.** In active learning, uncertainty estimation is used to determine whether or not a data point should be manually labeled and added to the training data. Question: Why can the disagreement (i.e. the difference in the prediction) of multiple neural networks be used to estimate the uncertainty of the prediction? Explain your answer briefly, if possible also with a sketch. **(2 points)**

3.1.7. **Dimensionality reduction.** You have 2-dimensional data points with coordinates x1 and x2. You want to reduce them to one dimension using principal component analysis and an autoencoder. Task: Draw a 2D point cloud where this is possible with both methods (without too much loss of information) and draw a point cloud where this is only possible with an autoencoder. Give brief reasons for your answers. **(2 points)**

Exam time was 90 minutes and there were german translations which I did copy.

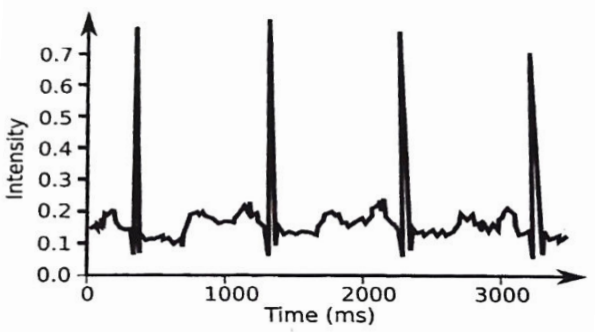
3.2. Machine learning applications (20 points in total)

3.2.1. **SMILES and generative models.** Draw the structure of the molecule proline from its SMILES string: O=C(O)C1CCCN1.
b) List and explain (briefly) the most important rules used for that example.
c) List one advantage and one disadvantage of using SMILES as output of generative models for molecular design. **(4 points)**

3.2.2. **Molecular fingerprints and generative models.** a) Explain the basic concept behind molecular fingerprints.
b) Would you use them as molecular representations in generative models for the design of molecules? Explain your answer. **(2 points)**

Exam time was 90 minutes and there were german translations which I did copy.

3.2.3. **Medical data.** You have a dataset with ECG data (heartbeat signals, i.e. intensity over time) labeled with 2 class labels (normal and not-normal). The dataset contains multiple time series with variable length and a fixed time interval of 1 ms, i.e. one intensity value per millisecond. Each time series belongs to one of the two classes, i.e. it contains hundreds of heartbeats (either all normal or all not-normal) which have an average length of approximately one second. Your task is to train a machine learning model to classify the ECG signals. You have the choice of recurrent neural network (RNN) and a convolutional neural network (CNN). Task: List one possible advantage and one possible disadvantage for each of both models, taking into account required data preprocessing steps. **(4 points)**



Exam time was 90 minutes and there were german translations which I did copy.

3.2.4. **Bayesian optimization.** A materials scientist in a lab has requested assistance from a colleague who is an expert in machine learning to design an algorithm to optimize the production parameters for a cathode used in electrical car batteries. The goal is to maximize the batteries capacity, measured as energy per kilogram (kWh/kg) of battery. The production parameters of the cathode to be optimized are the thickness of the electrode layer (in mm) and the processing temperature. Questions: a) Explain the steps of the basic workflow of Bayesian optimization in this situation.
b) Define the optimization parameters and the objective function.
c) Explain why Bayesian optimization is suitable for this task. **(3 points)**

Exam time was 90 minutes and there were german translations which I did copy.

3.2.5. **Graph neural networks.** Assume you have a dataset of molecules: Atoms with element types as nodes, chemical bonds as edges, and the geometry in form of cartesian coordinates of the atoms. You implement a graph neural network (GNN) with node and edge representations (i.e. vectors that describe nodes and vectors that describe edges) to be trained on the molecular dataset. You want to have a GNN model which is invariant toward translation and rotation of the input molecules. Invariant means that the model output is independent (i.e. does not change) of translations and rotations of the input.
Question: Do you use the geometry information in node or edge representations of the input to a GNN? How do you pre-process the cartesian coordinates to use them?
(2 points)

3.2.6. **Graph neural networks.** You want to train a generative model for molecules and have heard that graph neural networks (GNNs) are particularly well suited for regression and classification problems involving molecules. Now you design a variational autoencoder for molecules. As an encoder, you use a GNN with 3 message passing steps and a global aggregation step ("readout"), followed by a densely connected layer to generate the latent representation.
Question: Can you also use a GNN for the decoder? Explain your answer. **(2 points)**

Exam time was 90 minutes and there were german translations which I did copy.

3.2.7. **Molecular property prediction.** You have a large data set with 10,000 molecules (SMILES codes) and their toxicity (one scalar per molecule, i.e. SMILES code). You also have two weeks of access to the experiment used to generate the dataset and a large experimental molecular database of 100,000 untested and thus unlabeled molecules and their SMILES codes. Reliable toxicity prediction is necessary to find molecules that can be used as candidates for the development of new drugs, e.g. antibiotics. The experiments to test toxicity of molecules are too complex and expensive to be used for all molecules in the unlabeled database. 200 new molecules can be tested in parallel, which takes 24 hours overall. You are not sure if the 10,000 labeled molecules are representative of the unlabeled database.
Your task: Design and describe a machine learning based approach that finds the molecules with the lowest toxicity from the whole set of 110,000 molecules. What general workflow do you use? Which machine learning model do you use in that workflow? How are the molecules represented in this model? Justify your decisions. Use as much of the given information as possible. **(3 points)**