GENERATIVE HETERO-ENCODER MODEL FOR DE NOVO DESIGN OF SMALL-MOLECULE COMPOUNDS AS POTENTIAL INHIBITORS OF BCR-ABL TYROSINE KINASE

Karpenko Anna, Vaitko Timofey, Tuzikov Alexander, Andrianov Alexander
National Academy of Sciences of Belarus
Chronic myeloid leukemia (CML) is a clonal hematopoietic stem cell disorder and accounts for approximately 30% of the incidence of adult leukemias.

Currently available drugs have high toxicity and resistance.

The incidence of CML increases with age.


### Pipeline of solution

<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
<th>4. Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular libraries Molecular docking</td>
<td>Selection and development of neural network architecture Generate molecules</td>
<td>Molecular docking Properties prediction</td>
</tr>
<tr>
<td>Phase</td>
<td>1. Target discovery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Crystal structure of the ABL kinase domain associated with the DFG-out inhibitor AP24534

PDB ID: 3OXZ

Crystal structure of the domain of the mutant kinase ABL T315I associated with the DFG-out inhibitor AP24589

PDB ID: 3OY3
**Solution**

<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular docking</td>
</tr>
</tbody>
</table>

**PubChem**

120,000 compounds containing **aryl-aminopyrimidine**

**Molecular docking with rigid receptor and flexible ligand**

**Ligand-receptor binding energy** (Gibbs free energy)

**Feathers**

**SMILES**

**CANONICAL SMILES**

**String-based approach**
# Solution

<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
<td>Selection and development of neural network architecture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular docking</td>
<td>Generate molecules</td>
</tr>
</tbody>
</table>
Solution. ML WORKFLOW

<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
<td>Selection and development of neural network architecture</td>
</tr>
<tr>
<td></td>
<td>Molecular docking</td>
<td></td>
<td>Generate molecules</td>
</tr>
</tbody>
</table>

**Input Vectors**

- **X_S**: Smiles One Hot Encoded
- **X_CS**: Canonical Smiles One Hot Encoded
- **X_P**: Characteristics of molecules
- **X_E**: Binding Energy
- **X1_S**: Smiles One Hot Encoded
- **X1_CS**: Canonical Smiles One Hot Encoded
- **X1_P**: Characteristics of molecules

**Output Vectors**

- **G_S**: Smiles One Hot Encoded
- **G_CS**: Canonical Smiles One Hot Encoded

**Stages**

1. **LSTM ENCODER**
2. **DENSE ENCODER**
3. **CONCATENATION (LATENT SPACE)**
4. **DENSE DECODER**
5. **LSTM DECODER**

**Generated Vectors**

- **X_S[:-1]**
- **X_CS[:-1]**
- **G_S**
- **G_CS**

**Additional for generation:**

- **E**: desired energy
- **N**: random noise

**Optimization and Training**

- **Optimizer**: ADAM
- **Train size**: Epochs = 25
### Solution

<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
<th>4. Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
<td>Selection and development of neural network architecture</td>
<td>Molecular docking</td>
</tr>
<tr>
<td></td>
<td>Molecular docking</td>
<td>Generate molecules</td>
<td></td>
<td>Properties prediction</td>
</tr>
</tbody>
</table>

\[
LF(s) = CCE(s) + 0.1 \cdot CCL(s),
\]

- \(CCE(s)\) is the categorical cross entropy,
- \(s\) is a molecule in the SMILES format,
- \(CCL(s)\) (CustomChemLoss) is the function that imposes penalties for violations of a molecule stereochemistry and the absence of 2-arylaminopyrimidine in its chemical structure.
<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
<th>4. Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
<td>Selection and development of neural network architecture</td>
<td>Molecular docking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular docking</td>
<td>Generate molecules</td>
<td>Properties prediction</td>
</tr>
</tbody>
</table>

1083 unique molecules have been generated
<table>
<thead>
<tr>
<th>Phase</th>
<th>1. Target discovery</th>
<th>2. Screening</th>
<th>3. Lead generation</th>
<th>4. Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Find all targets from literature and Protein Data Bank</td>
<td>Create a molecular library</td>
<td>Selection and development of neural network architecture</td>
<td>Molecular docking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular docking</td>
<td>Generate molecules</td>
<td>Properties prediction</td>
</tr>
</tbody>
</table>

\[ ECR = \sum_{sf} \frac{1}{\sigma_{sf}} \cdot \exp\left(-\frac{\text{rank}_{sf}}{\sigma_{sf}}\right), \]
Results

CrossECR = 0.0674

C22H22N6

CrossECR = 0.0835

C25H34N8O

CrossECR = 0.0674

C30H27N7O2

CrossECR = 0.0931

C20H19ClN6O

Ponatinib

CrossECR = 0.0399

\[
crossECR(i) = \frac{ECR_1(i)}{\max_i\{ECR_1(i)\}} + \frac{ECR_2(i)}{\max_i\{ECR_2(i)\}},
\]
THANK YOU

This study was supported by the State Program of Scientific Research “Convergence 2025” (subprogram “Interdisciplinary research and emerging technologies”, project 3.4.1).

Hanna Karpenka
rfe.Karpenko@gmail.com
https://www.linkedin.com/in/anna-karpenko-by/