Insights from Lead Optimization Efforts Using KNIME in Industry

Thomas M Kaiser, BM BCh, PhD
Chief Scientific Officer
Avicenna Biosciences
Who We Are

• Avicenna Biosciences is first and foremost a drug development firm that generates NCEs using medicinal chemistry and machine learning

• Every machine learning scientist in Avicenna trained as either a chemist or a physicist first

• We work exclusively on solving DMPK/Tox problems to enable quality chemical matter for innovative clinical trials

• Launched in 2019, we now have multiple programs in Oncology, Neurodegeneration/Neuroinflammation and Autoimmune/Autoinflammatory indications

• Future work will move us from purely development problems to more discovery-type programs through our work on dataset augmentation with physics-based methods
Some Difficulties in Applying ML to Drug Development

• Addressing a true drug development need is a major problem – the translation of a medicinal chemistry design point to a machine learning experiment has been a major hurdle, and the clarity of machine learning experimental design has been low in the past

• As an example, there is a miscommunication between the medicinal chemists discussing multiobjective optimization and the ML people who hear “end-to-end”

• Additionally, the process of data sourcing and curation has limited transparency and no established process for formal presentation either to internal or external audiences

• We have developed two tools that aid us in designing algorithms for our internal programs: ML experiment design diagrams and Schematic of Literature Inclusion Criteria for Experiment in ML (SLICE ML)
ML Experimental Design

• The applicability domain for various ML methods is not equivalent for all methods, and some methods have limited utility for problems within chemical biology and drug development/discovery

• In our experience, there is a communication gulf between machine learning scientists and medicinal chemists/pharmacologists

• This miscommunication can result in the selection of ML methods which fail to have utility for predicting desired solutions to discovery or development problems

• A way of representing the design of machine learning experiments that is accessible to non-ML scientists would reduce miscommunication
ML Experimental Design

<table>
<thead>
<tr>
<th>Algorithm Type</th>
<th>Random Forrest (Tree Conditions: Information Gain Ratio, Limited Tree Depth &lt; 15, No Node Size Minimum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trees</td>
<td>200</td>
</tr>
<tr>
<td>Learning Type</td>
<td>Supervised Learning</td>
</tr>
<tr>
<td>Point of Run Replication</td>
<td>Test/Train Split</td>
</tr>
<tr>
<td>Number of Replicate Runs</td>
<td>Triplicate</td>
</tr>
<tr>
<td>Independent Variable</td>
<td>ECFP4</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>Active = (1,0)</td>
</tr>
</tbody>
</table>
FEPML Background – Theory

• Machine learning in combination with Relative Binding Free Energy (RBFE) calculations
  • Machine learnings applicability domain is limited to the availability of data
  • How do we overcome the limitations of information poor projects?
  • RBFE has emerged as highly accurate molecular mechanics methods to predict binding affinity of similar compounds to a given target (1-2kcal/mol)
    • FEP is currently the gold standard

• Rationale
  • FEP calculations can serve as an input to ML algorithms to partially overcome information sparse limitations
  • Reduce time and cost associated of traditional medicinal chemistry efforts ($100-150 vs $2000-5000)

ML Experimental Design Diagrams

1. **cSrc IC_{50} Set n = 133**
   - Calculate ECFP4 Fingerprints for n = 133
   - **Training Set for Information-Poor Algorithm n = 11**

2. **cSrc IC_{50} - 11 Initial Set n = 122**
   - Split for < Supervised Learning Active Threshold
   - **Active**
   - **Inactive**

3. **Total Training n = 11**
   - **Active**
   - **Inactive**

4. **Random Forest**
   - **Prediction**

5. **60/40 Train/Test Split**
   - Data with Real IC_{50} Added to Training Set n = 73
   - Test Set n = 49
ML Experimental Design Diagrams

Kaiser, T. M.; Burger, P. B., unpublished
### ML Experimental Design Table

<table>
<thead>
<tr>
<th>Algorithm Type</th>
<th>Random Forest (Tree Conditions: Gini Split Criterion, No Maximum Tree Depth, No Node Size Minimum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trees</td>
<td>1000</td>
</tr>
<tr>
<td>Learning Type</td>
<td>Supervised Learning</td>
</tr>
<tr>
<td>Point of Run Replication</td>
<td>n = 11/122 partitioning</td>
</tr>
<tr>
<td>Number of Replicate Runs</td>
<td>10-fold</td>
</tr>
<tr>
<td>Independent Variable</td>
<td>ECFP4</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>Active = (1,0)</td>
</tr>
</tbody>
</table>

Kaiser, T. M.; Burger, P. B., *unpublished*
Lessons from Systematic Review and Meta-Analysis

- Machine learning involving multiple sets of literature and intra-organizational data is inherently a form of meta-analysis.

- Medicine has explored solutions for transparency issues in experimental design for meta-analysis.

- The solution most commonly employed is the use of the systematic rigor of inclusion/exclusion of data provided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).
Schematic of Literature Inclusion Criteria for Experiments in Machine Learning - SLICE ML

Data Sources

- Records identified from: External Databases (n = )
- Internal Databases (n = )

Records removed before Screening:
- Duplicate records removed (n = )
- Records removed for other reasons (n = )

Dataset Cleaning

- Records screened (n = )

Datapoint removed by General Screening:
- Records with unknown assay type (n = )
- Records with missing values (n = )
- Records with missing structure (n = )

Datapoints assessed for eligibility in for ML Experiment (n = )

Datapoint removed by Specific Exclusion Criteria:
- Reason 1 (n = ) [e.g. data value outside typical range]
- Reason 2 (n = ) [e.g. data value generated at edge of dose-response curve]
- Reason 3 (n = ) [e.g. data value is for non-relevant assay type] etc.

Final Dataset (n = )
Conclusions

• We have drawn on other disciplines to generate methods for a rigorous standardization that allows machine learning, chemistry and biology to integrate into a single environment

• Clear diagrams of the machine learning experiment have enabled better translation of chemical or biological information into machine learning systems

• The formalization and transparent representation of the process of data cleaning for ML through SLICE ML has enabled more robust applications in our drug development process