Artificial intelligence-driven design of fuel mixtures: Supplementary material

Supplementary Note 1

Over the last decade, DL outperformed AI technologies based on other ML algorithms, the advantage of DL is the flexibility of neural network architecture, which can be tailor-made for the specific task. It achieves great success by learning to represent input as a nested hierarchy of concepts, with more abstract representations derived from less abstract ones [1]. DNN parameters are trained to transform original data into another representation which is a vector with the most semantic features for the prediction of joint properties, thus in this respect, DL is a form of feature learning based on the non-linear mappings. This intermediate vector is referred to as a latent space or hidden space representation, which is continuous and differentiable [2]. For a task of designing fuel matching not one, but multiple combustion-related properties, it is crucial that the latent vector of each datum is consistent and sufficiently informative to predict all properties of interest. Merging databases of different properties helps to fill in gaps in hidden space, thus, widens the chemical space, however it may also arise issue with respect to missing data.

Prior works have been reported to apply ML methods to predict multiple combustion-related properties of single components. For example, Schweidtmann et al. [3] proposed Graph Neural Network (GNN) that maps molecular graph to the three ignition quality indicators, the Derived Cetane Number (DCN), RON and MON. A single neural network model was trained to predict multiple thermochemical properties and used to screen high energy components [4]. These models can be used to screen a wide variety of hydrocarbon fuels, but its applicability cannot be extended to screening on a mixture level.

The prediction of mixture properties remains also a challenge of great importance in the inverse fuel design since the composition of fuels can be complex. A simple linear-mixing-by-mole rule was proposed by Knop et al. [5] to estimate octane numbers based on values of pure species and results demonstrated high accuracy of RON and MON predictions for n-heptane/iso-octane/toluene mixtures. However, there is a growing evidence that octane enhancers, such as some oxygenates, do not follow a linear mixing rule and can have both synergistic or antagonistic effect on overall ignition quality depending on the composition and properties of the base fuel and the blendstock. These studied octane boosters include bioethanol [6-10], prenol [11], terpineol [12], furan mixtures, i.e. methylfuran and dimethylfuran] [13, 14].

In addition, mixture properties are currently evaluated through mixing rules at the descriptor [15] or physical properties [16] level. Alternatively, a lot of research effort has been invested with the goal finding the most informative representation for mixtures, which can be directly used as an input without incorporation of mixing rules. For example, the input based on measurements
obtained via various analytical techniques was shown to be efficient in predicting ONs of complex mixtures, these techniques include flame spectroscopy emission [17], Near-Infrared (NIR) Spectroscopy [18], gas chromatography [19], nuclear magnetic resonance (NMR) spectroscopy [20, 21], ignition quality tester (IQT) [22]. Data from these methods were mapped into RON and MON using partial least-squares [17, 18], multilinear regression (MLR) [20], artificial neural networks [21, 23]. Recent joint Birmingham-Shenzhen work [24] featured a machine learning approach to predict 15 different properties, mixtures properties being obtained through functional group interaction. The mixture representations used in these models cannot be directly connected to the representations of pure components, thus, the configuration of these models is ineligible for inverse fuel design.

![Fig. S1: Backward fuel design workflow.](image)

**Table S1:** Performance of the proposed model on the test set.

<table>
<thead>
<tr>
<th>Property</th>
<th>$R^2$</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pure</td>
<td>0.89</td>
<td>4.73</td>
</tr>
<tr>
<td>mix</td>
<td>0.98</td>
<td>1.76</td>
</tr>
<tr>
<td>all</td>
<td>0.93</td>
<td>3.05</td>
</tr>
<tr>
<td>MON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pure</td>
<td>0.89</td>
<td>3.85</td>
</tr>
<tr>
<td>mix</td>
<td>0.96</td>
<td>1.85</td>
</tr>
<tr>
<td>all</td>
<td>0.92</td>
<td>2.8</td>
</tr>
<tr>
<td>YSI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pure</td>
<td>0.99</td>
<td>4.58</td>
</tr>
<tr>
<td>mix</td>
<td>0.95</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Fig. S2: Pie charts of final 4 candidates with predicted RON/MON/YSI value and mean squared error from the target properties. **a** mixture 22-13 with properties 94.9/85.3/59.6 (mse = 0.07), **b** mixture 18-17: 95.4/85.1/59.6 (mse = 0.09), **c** mixture 26-20: 94.7/85.4/59.8 (mse = 0.09).
**Fig. S2**: Pie charts of final 4 candidates with predicted RON/MON/YSI value and mean squared error from the target properties (cont.). *d* mixture 6-1: 95/85/60 (mse = 0.00), *e* mixture 17-1: 95/85/60 (mse = 0.00).

**Table S2**: Supplementary properties of new potential components from short-listed mixtures.

<table>
<thead>
<tr>
<th>Compound</th>
<th>FKMC</th>
<th>EPIB7</th>
<th>WS ($g L^{-1}$)</th>
<th>AHL (d)</th>
<th>MP (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tetramethoxy methane</td>
<td>yes</td>
<td>-0.845</td>
<td>110</td>
<td>3.2</td>
<td>-5</td>
</tr>
<tr>
<td>Isopentyl acetate</td>
<td>yes</td>
<td>0.655</td>
<td>2.0</td>
<td>1.8</td>
<td>-78</td>
</tr>
<tr>
<td>Ethylidene cyclopentane</td>
<td>yes</td>
<td>0.203</td>
<td>0.1</td>
<td>0.12</td>
<td>-129</td>
</tr>
</tbody>
</table>

*Notation: FKMC = FKM compatibility, EPIB7 = EPI Biowin 7, WS = Water solubility, AHL = Atmospheric half-life, MP = Melting point.*
### Table S3: Numerical scales in YSI measurements.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Lower endpoint</th>
<th>Upper endpoint</th>
<th>YSI&lt;sub&gt;A&lt;/sub&gt;</th>
<th>YSI&lt;sub&gt;B&lt;/sub&gt;</th>
<th>Species</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-heptane</td>
<td>tetralin</td>
<td>10</td>
<td>100</td>
<td></td>
<td>diesel surrogates</td>
<td>[25]</td>
</tr>
<tr>
<td>low soot</td>
<td>n-hexane</td>
<td>benzene</td>
<td>0</td>
<td>100</td>
<td>pure species, diesel and jet fuel surrogates</td>
<td>[26, 27] [28]</td>
</tr>
<tr>
<td>high soot</td>
<td>benzene</td>
<td>1,2-dihydro-naphthalene</td>
<td>30</td>
<td>100</td>
<td>pure species</td>
<td>[29, 30]</td>
</tr>
<tr>
<td></td>
<td>benzene</td>
<td>toluene</td>
<td>30</td>
<td>43.5</td>
<td>gasoline surrogates</td>
<td>[31]</td>
</tr>
<tr>
<td>unified</td>
<td>n-hexane</td>
<td>benzene</td>
<td>30</td>
<td>100</td>
<td>pure species</td>
<td>[28]</td>
</tr>
<tr>
<td></td>
<td>n-heptane</td>
<td>toluene</td>
<td>36</td>
<td>170.9</td>
<td>TG and surrogates</td>
<td>[32]</td>
</tr>
</tbody>
</table>

### Table S4: Train/test split with hierarchical stratified sampling.

<table>
<thead>
<tr>
<th>Subset #</th>
<th>strata</th>
<th>train</th>
<th>test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub 1</td>
<td>pure</td>
<td>115</td>
<td>21</td>
<td>136</td>
</tr>
<tr>
<td>RON, MON and YSI</td>
<td>mix</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sub 2</td>
<td>pure</td>
<td>159</td>
<td>28</td>
<td>187</td>
</tr>
<tr>
<td>RON and MON</td>
<td>mix</td>
<td>294</td>
<td>53</td>
<td>347</td>
</tr>
<tr>
<td>Sub 3</td>
<td>pure</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>RON and YSI</td>
<td>mix</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sub 4</td>
<td>pure</td>
<td>23</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>RON</td>
<td>mix</td>
<td>81</td>
<td>15</td>
<td>96</td>
</tr>
<tr>
<td>Sub 5</td>
<td>pure</td>
<td>260</td>
<td>47</td>
<td>307</td>
</tr>
<tr>
<td>YSI</td>
<td>mix</td>
<td>30</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Sub 6</td>
<td>pure</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MON</td>
<td>mix</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

### Table S5: Tuned hyperparameters.

<table>
<thead>
<tr>
<th>Net</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Optimizer</td>
<td>ADAM</td>
</tr>
<tr>
<td></td>
<td>Batch size</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Latent vector dim (m)</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td># LSTM layers</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Extractor 1 hidden states dim</td>
<td>[256, 64, 16]</td>
</tr>
<tr>
<td></td>
<td>nodes dim in FC layer</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Extractor 2 # FC layers in</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>nodes dim</td>
<td>[1048, 512, 12]</td>
</tr>
<tr>
<td></td>
<td>Predictive # FC layers</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>nodes dim</td>
<td>[128, 24, 12]</td>
</tr>
</tbody>
</table>
**Fig. S3:** Distribution of target values in database. Histograms shown for a RON and MON, and b YSI.

**Fig. S4:** Depth-first search algorithm.
Table S6: Gasoline specifications used as linear constraints.

<table>
<thead>
<tr>
<th>Species category</th>
<th>Upper threshold (% v/v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromatics</td>
<td>35</td>
</tr>
<tr>
<td>Benzene</td>
<td>1</td>
</tr>
<tr>
<td>Olefins</td>
<td>18</td>
</tr>
<tr>
<td>Diolefins</td>
<td>1</td>
</tr>
<tr>
<td>Ethers</td>
<td>22</td>
</tr>
<tr>
<td>Cyclic ethers, furans and hydrofurans</td>
<td>1</td>
</tr>
<tr>
<td>Ethanol</td>
<td>10</td>
</tr>
<tr>
<td>Methanol</td>
<td>3</td>
</tr>
<tr>
<td>Isopropanol</td>
<td>12</td>
</tr>
<tr>
<td>Butanol</td>
<td>15</td>
</tr>
<tr>
<td>Other oxygenates</td>
<td>15</td>
</tr>
<tr>
<td>All Oxygenates</td>
<td>30</td>
</tr>
<tr>
<td>C\textsubscript{3}, C\textsubscript{4}, C\textsubscript{7} and C\textsubscript{8} aliphatic rings</td>
<td>1</td>
</tr>
</tbody>
</table>
Supplementary References


Kashif, M., Bonnety, J., Matynia, A., Da Costa, P. & Legros, G. Sooting propensities of some gasoline surrogate fuels: Combined effects of fuel blending and air vitiation. *Combustion and Flame*