Application of Artificial Intelligence for the Prediction of Solvation Free Energies for Covid-19 Drug Discovery

Sampreeth Immidisetty[1] and Deepak Agrawal[#]

[1]Greenwood High International School, Banglore, Karnataka, India
[#]Advisor

ABSTRACT

Solvation free energy is a key indicator of the effectiveness of a drug molecule. There are several applications of predicting the solvation free energies of chemical compounds using quantum mechanical methods. However, these methods take a long time and are costly. For that reason, the application of recently developed artificial intelligence techniques for the prediction of solvation free energies is becoming increasingly valuable in drug discovery to address time and the high-cost issues with traditional quantum mechanical approaches. In this paper, we present application of two different artificial intelligence models for predicting solute-solvent free solvation energy for Covid-19 drug design. The research involves building, training, evaluating and comparing the performances of the two models on a large dataset, then predicting solvation free energies for 138 known APIs and 28 organic solvents that could potentially be used as a Covid-19 medicine. The potential repurposing of 138 drugs for Covid-19 from solubility perspective is novel. We demonstrate the application of the AI models and derive several conclusions regarding suitability of the APIs and their efficacy. We conclude our research by providing insights on how our work can be put to future use towards drug development.

Introduction

The search for an effective treatment for Covid-19 disease is an ongoing global research effort. Our research seeks to contribute to that effort by systematically estimating the solvation energy of a number of known drug APIs paired with a number of known organic solvents. The objectives are two-fold, one to demonstrate use of Artificial Intelligence (AI) models in drug design, and second, to find promising pairs of Active Pharmaceutical Ingredients (APIs) and organic solvents to treat Covid-19 disease. This effort to find an effective repurposed drug for Covid-19 among several known APIs (or solutes) is novel which directly contributes to the knowledge regarding potential cure for Covid-19.

The solute-solvent pairing to obtain maximum solubility is an important requirement in drug design, because solubility helps in drug absorption and retention which in turn determines the bioavailability of the API, i.e., portion available for action against the target. It is a tedious task to compute solubility in a lab experimentally (in-vitro); for that reason, reliable computational models (in-silico) for predicting the solubility of an API against several solvents are sought-after in drug discovery because in-silico models can assess the most promising solvents for a given API in a much shorter amount of time. The in-silico model outcomes can then be tested further in the laboratory (in-vitro) to further close-in on the most promising solute-solvent pairs for further in-vivo studies.

The metric for solubility is solvation free energy; we optimize solvation free energy to obtain the most promising solute-solvent pairs. Solvation energy is the amount of energy generated when a solute is dissolved in a solvent. A negative solvation energy is associated with an exothermic reaction, whereas a positive solvation energy is associated with an endothermic reaction.

Typically, the solvation energy ranges from -5 to -20, where the smaller the solvation energy (meaning larger negatives) implies more solubility and therefore better solute-solvent pairing.

Previously, quantum mechanical methods used for the prediction of solvation energy (Duarte Ramos Matos et al., 2017, Kröger et al., 2020) generally involve high computational costs and time; and for that reason, Artificial Intelligence methods are being increasingly utilized for the prediction of solvation energy. In this research we implement two different but competing Artificial Intelligence models to estimate the solvation energy of several solute-solvent pairs. The two models are described next.

CIGIN2 model

CIGIN2 model (Pathak et al., 2020), an acronym for Chemically Interpretable Graph Interaction Network, is a graph neural network model to predict solvation free energies. There are 3 parts in CIGIN2 model: message passing (Gilmer et al., 2017), interaction mapping, and prediction phase.

The first phase, Message passing, is a neural network in which inter-atomic interactions are computed within solute and solvent molecules represented as molecular graphs. A molecule represented as a graph has the atoms represented as the nodes and the bonds as the edges. Both the nodes and the edges are characterized by a set of features. These feature vectors are constantly updated over a certain amount of time steps based on their environment (specifically the neighboring nodes). The final feature vectors are obtained by gathering layers.

The second phase involves computation of a solute-solvent interaction map which captures the electronic and steric factors that govern the solubility of molecules. This interaction map can provide useful insights on different solute and solvents' features impact on solvation free energy.

And the third phase is about prediction of solvation free energies using solute-solvent interaction maps and features from the message passing phase. This involves passing the outputs through the set2set readout layer (Vinyals, Bengio, and Kudlur 2016), followed by a fully connected multi-layer perceptron, using the rectifier unit activation (ReLU) function, with an output layer containing the final solvation free energy predictions.

CIGIN2 model is trained using MNSOL data of solvation energies of 2049 pairs of 418 solutes and 91 solvents, a validation dataset of 228 pairs, and a test dataset of 253 pairs.

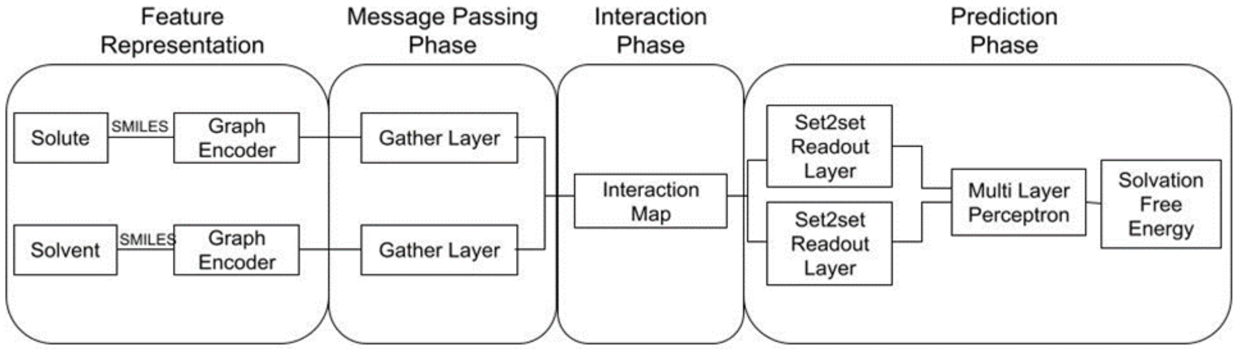

Figure 1. CIGIN2 Model Architecture

DELFOS model

DELFOS model (Lim et al., 2019), a Deep Learning model for solvation free energies in generic organic solvents, is a Quantitative Structure–Property Relationship (QSPR) method which predicts solvation free energy of organic solute and solvents from their empirical or structural features.

DELFOS model uses two separate solvent and solute encoder networks (sub-neural networks). The primary architecture of the encoder is based on two bidirectional recurrent neural networks. The encoders first embed the chemical structure of the given solute and solvent into a molecular descriptor using Mol2Vec word embedding model - where an atom or a substructure is a word and a molecule is a sentence (Jaeger et al., 2018, Pennington et al., 2014). The Mol2Vec word embedder uses the Morgan algorithm to generate substructure vectors for each atom based on their environment (Morgan 1965). Then, the encoder uses a bi-directional RNN layer (Schuster et al., 1997); augmented with a dot shared dot product attention layer to extract important sub-structures from outputs of recurrent neural networks (Bahdanau et al., 2014). The interaction between the hidden states in the shared attention layer can offer information about which sub-structures play a dominant role in the solvation process.

Finally, the 3rd sub-neural network, the mapping function or the predictor neural network has a single fully connected perceptron layer with a rectifier unit (ReLU) and an output layer. It uses the concatenated feature of the solvent and solute [u; v] as an input. The predictor neural network with a fully connected MLP layer calculates the solvation free energy of a given solvent–solute pair using the feature vectors from the two encoders.

DELFOS model is also trained using MNSOL data of solvation energies of the same set of 2049 pairs of 418 solutes and 91 solvents as for CIGIN2. And the same validation dataset of 228 pairs, and test dataset of 253 pairs.

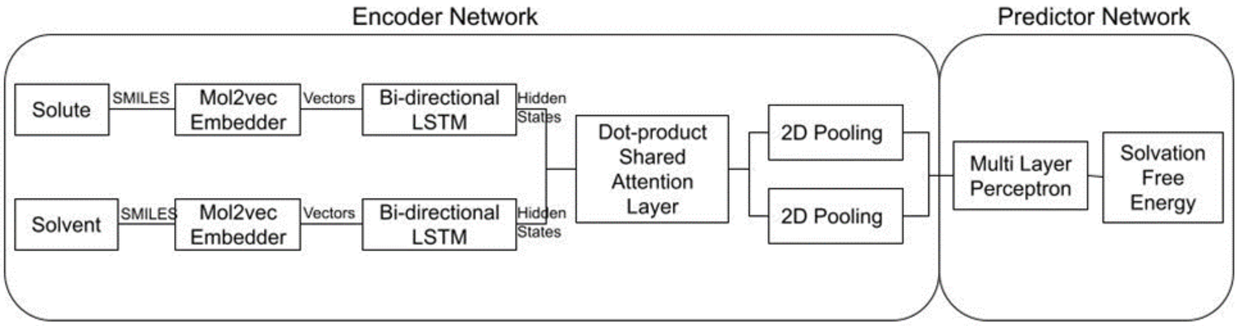

Figure 2. DELFOS Model Architecture

List of Solutes and Solvents

In this research, we compiled a list of 138 solutes based on extensive literature review of known APIs that could treat Covid-19 disease. These APIs are listed in Table 1. Using existing known APIs to treat a different disease, different from what the API was originally developed to treat, is known in the literature as Drug Re-purposing or "DRP" effort. DRP for Covid-19 is being investigated the world over and this research contributes to DRP efforts by examining a new list of known APIs from solubility perspective as a Covid-19 medicine.

We test solubility of 138 known APIs against a list of 28 known organic solvents. We compiled solvents list based on extensive literature review of the known solvents. These solvents are listed in Table 2.

Table 1. List of 138 Known APIs

A. Alpha amino acids	**B. Anaplastic lymphoma kinase**	**C. Anthelmintics**
1 CC-223	2 LDK378	3 Ivermectin
D. Antiarrhythmic	**E. Antibacterial**	**F. Antibiotic**
4 Dronedarone HCl	5 Hexachlorophene, 6 Sulfadoxine	7 Azithromycine
G. Anticancer	**H. Anticholinergics**	**I. Antidepressants**

8 Abemaciclib, 9 Amuvatinib, 10 Carboxyamidotriazole, 11 Gilteritinib, 12 GSK2606414, 13 Homoharringtonine, 14 Imatinib Mesylate, 15 LDE225, 16 LGK-974, 17 LY2228820, 18 Osimertinib mesylate, 19 Pevonedistat, 20 Pexidartinib, 21 Regorafenib, 22 Sorafenib, 23 Tamoxifen Citrate, 24 Tyrphostin, 25 Vatalanib	26 Benztropine Mesylate	27 Clomipramine Hydrochloride
J. Antidiarrheal	**K. Antifibrotic**	**L. Antifungal**
28 Loperamide	29 PF-670462	30 Cloconazole, 31 Oxiconazole, 32 Ravuconazole, 33 Chlormidazole, 34 Ketoconazole
M. Antihistamine	**N. Antihypertensive**	**O. Antiinflammatory**
35 Clemizole hydrochloride, 36 Mequitazine, 37 Loratadine 38 Ebastine	39 Berbamine hydrochloride	40 PH-797804, 41 CVL218
P. Antileukemia	**Q. Antilipemic**	**R. Antimalarial**
42 Tioguanine, 43 Alvocidib, 44 AI-10-4	45 Triparanol	46 Quinacrine hydrochloride monohydrate, 47 Mefloquine Hydrochloride, 48 Amodiaquin Dihydrochloride Dihydrate, 49 Amodiaquine hydrochloride, 50 Amodiaquin Hydrochloride, 51 Chloroquine, 52 Chloroquine Phosphate, 53 Hydroxychloroquine Sulfate
S. Antimethemoglobinemia	**T. Antiparasitic**	**U. Antiproliferative**
54 Methylene blue	55 Oxyclozanide	C1NDSS5
V. Antiprotozoal	**W. Antipsychotic**	**X. Antiretroviral**
57 Emetine	58 Chlorpromazine Hydrochloride, 59 Fluspirilene, 60 Penfluridol, 61 Thioridazine hydrochloride, 62 CBIPES	63 Amprenavir, 64 Atazanavir, 65 Dapivirine, 66 Dolutegravir, 67 Indinavir, 68 Lopinavir, 69 Nelfinavir, 70 Saquinavir, 71 Tipranavir
Y. Antiseptic	**Z. Antispasmodic**	**AA. Antitapeworm**
72 Cetylpyridinium chloride, 73 Octenidine	74 Drotaverine	75 Niclosamide
AB. Antitumor	**AC. Antiviral**	**AD. Anxiolytic**
76 IPAG, 77 Tetrandrine	78 Arbidol, 79 Darunavir, 80 Favipiravir, 81 Penciclovir, 82 Remdesivir, 83 Ribavirin, 84 Nitazoxanide, 85 Tilorone, 86 Harringtonine	87 ZK-93423, 88 Opipramol dihydrochloride, 89 Etifoxine
AE. Atypical antipsychotics	**AF. Beta blockers**	**AG. Bioactive isoflavone**
90 Adoprazine, 91 Brexpiprazole	92 Oxprenolol hydrochloride	93 Osajin
AH. Biphenyls	**AI. Bisbenzylisoquinoline alkaloids**	**AJ. Calcium entry blocker**

94 Mibampator	95 Cepharanthine	96 Flunarizine
AK. Capsaicin-induced antihyperalgesia	**AL. Chronic pancreatitis**	**AM. Coronary vasodilator**
97 AMG-9810	98 Camostat	99 Lidoflazine
AN. Corticosteroids	**AO. Cystic fibrosis**	**AP. Dopamine antagonist**
100 Ciclesonide, 101 Loteprednol etabonate	102 Ivacaftor	103 Thiethylperazine Maleate
AQ. Dopamine D3 receptor	**AR. Estrogen agonist**	**AS. Estrogen receptor**
104 BP-897	105 Bazedoxifene, 106 Toremifene Citrate	107 Droloxifene((E)-3-Hydroxy tamoxifen)
AT. Heart Treatment	**AU. Hutchinson-Gilford progeria syndrome**	**AV. Immunomodulators**
108 Digoxin, 109 Lanatoside C, 110 Ouabain	111 Lonafarnib	112 JTE-013
AW. Immunosuppressant	**AX. Multidrug-resistant cancer cells**	**AY. Opioid receptor**
113 Cyclosporine	114 Isoosajin, 115 Isopomiferin	116 SB-612111
AZ. Opium alkaloid antispasmodic	**BA. Oral urinary analgesic**	**BB. Oral anticholesterol**
117 Papaverine	118 Phenazopyridine	119 Asimibe
BC. Ovulatory stimulant	**BD. Peripheral vasodilator**	**BE. Phenothiazines**
120 Clomiphene Citrate	121 Ethaverine	122 Fluphenazine Dihydrochloride, 123 Promethazine Hydrochloride
BF. Photodynamic Therapy	**BG. Platelet thrombopoietin receptor**	**BH. Progestin**
124 Hematoporphyrin	125 Avatrombopag	126 Hydroxyprogesterone caproate
BI. Prophylactic antianginal	**BJ. Proton-pump inhibitors**	**BK. Purine analogue**
127 Perhexiline maleate	128 Omeprazole	129 Thioguanosine
BL. Pyranoxanthones	**BM. Respiratory stimulant**	**BN. Sclerosing agent**
130 Dihydrogambogic acid	131 Almitrine	132 Polidocanol
BO. Synthetic organoselenium	**BP. Thrombocytopenia**	**BQ. Thrombopoietin receptor agonists**
133 Ebselen	134 Eltrombopag	135 Lusutrombopag
BR. Treatment of fungal infection	**BS. Vasodilators**	**BT. VRAC inhibitor**
136 Terconazole Vetranal	137 Alprostadil	138 DCPIB

Table 2. List of 28 known Organic Solvents

1 Acetic acid	2 Heptane	3 Acetone	4 Isobutyl acetate
5 Anisole	6 Isopropyl acetate	7 1-Butanol	8 Methyl acetate
9 2-Butanol	10 3-Methyl-1-butanol	11 Butyl acetate	12 Methylethyl ketone
13 tert-Butylmethyl ether	14 2-Methyl-1-propanol	15 Dimethyl sulfoxide	16 Pentane
17 Ethanol	18 1-Pentanol	19 Ethyl acetate	20 1-Propanol
21 Ethyl ether	22 2-Propanol	23 Ethyl formate	24 Propyl acetate
25 Formic acid	26 Triethylamin	27 Water	28 N, N-Dimethylformamide

We use two AI models CIGIN2 and DELFOS to predict solvation energy for each of the 138 APIs with each of the 28 solvents, with the objective of discovering most promising solute-solvent pairs on the basis of lowest solvation energy.

The predictions of solute-solvent solvation energies are presented in Figure 3 from CIGIN2 model and in Figure 4 from DELFOS model.

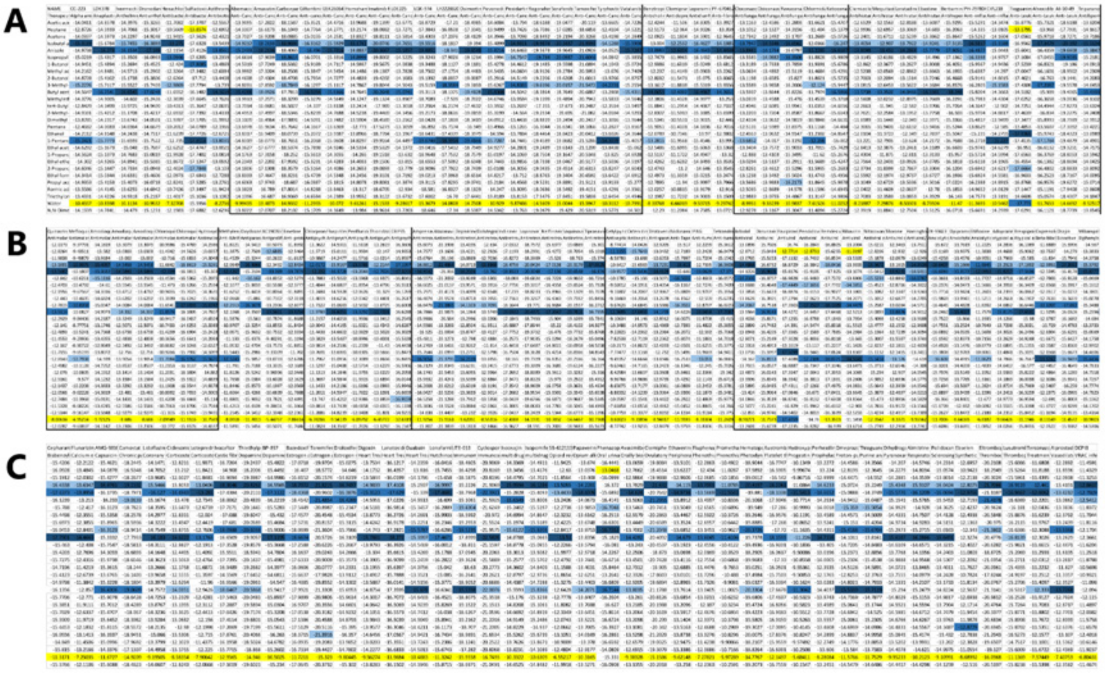

Figure 3. CIGIN2 Model Solvation Energy Predictions for 138 API Solutes x 28 Solvents

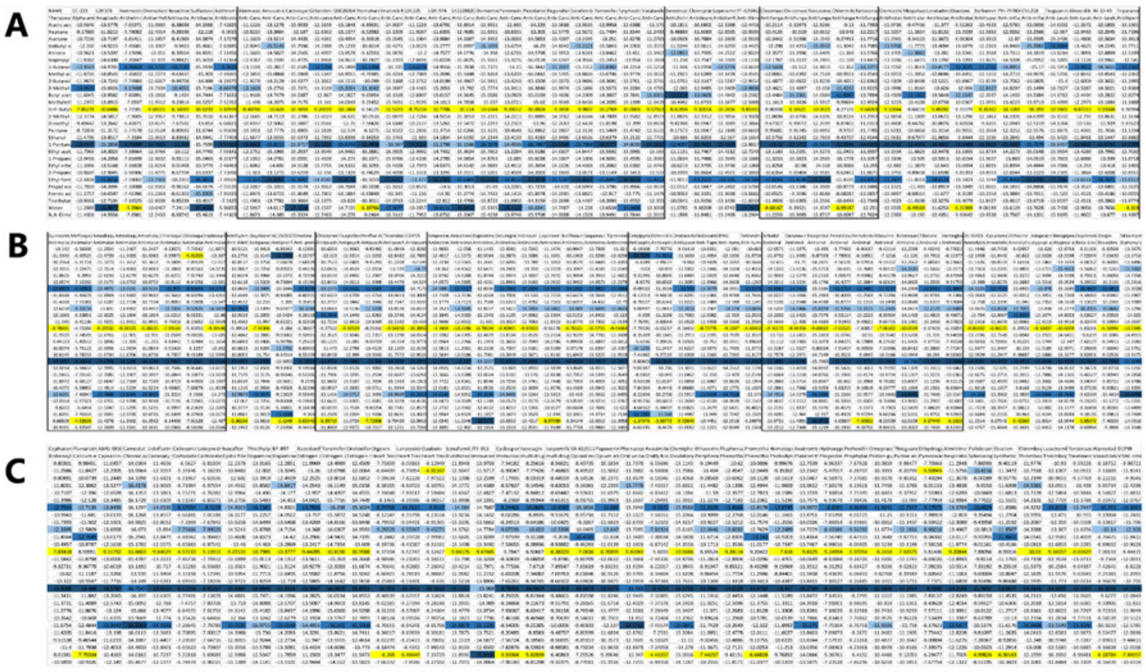

Figure 4. DELFOS Model Solvation Energy Predictions for 138 API Solutes x 28 Solvents

MNSOL Dataset

In this research, the CIGIN2 and DELFOS models are trained using The Minnesota Solvation Database, created by the Department of Chemistry, Chemical Theory Centre, and Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA. The database consists of a total of 3037 solvation free energy data points. This composes of experimental aqueous solvation free energies of 274 neutral solutes, 31 clustered ions consisting of a single water molecule and 112 ionic solutes. Additionally, it consists of 87 solvation free energies of water and 11 organic solvents for 64 neutral solutes as well as 2002 solvation free energies between 322 neutral solutes in 90 organic solvents. The solutes at most consist of the element's H, C, N, O, F, Si, P, S, Cl, and Br.

In total, MNSOL has 3037 experimental free solvation energies of 790 unique solutes in 92 solvents (Marenich et al., 2012) including neutral solutes, charged solutes, and samples of transfer free energies, which after removal of charged solutes and sample of transfer free energies results in a final usable dataset of 2530 unique combinations of solute and solvent in this paper.

Model Training/ Validation/ Testing

Both AI models CIGIN2 and DELFOS in this research utilize the same MNSOL dataset of solvation energies and the exact same training, validation and test datasets. The final usable MNSOL dataset containing 2530 data points was split as follows: 80% train (2049 data points), 10% validation (228 data points), and 10% test (253 data points).

The models were trained on 80 epochs with a batch size of 32 on the training dataset. The model states with the lowest MAE score on the validation set were saved and used to evaluate each models' performance on the test set using 3 evaluation metrics:

1. Mean Absolute Error: Calculated as the average of the absolute values of the residuals
2. Root Mean Squared Error - Calculated as the square root of the average of the square of the residuals

3. Mean Absolute Percentage Error - Calculated as the average of the absolute values of the percentages of the residuals.

The code for both models was written in Python. CIGIN2 was executed using PyTorch, while DELFOS was designed and executed using PyTorch as well as Tensorflow. The PyTorch based models yield the following performance metrics on the test set

Table 3. CIGIN2 and DELFOS test set performance metrics and benchmark scores.

Model	MAE	RMSE	MAPE	Benchmark* (RMSE)
CIGIN2	0.5323	1.178	28.21	RMSE: 0.57**
DELFOS	0.5615	1.304	36.64	RMSE: 0.57**

*Note: Benchmark models' RMSE mentioned here is from the respective research papers of CIGIN2 and DELFOS.
**Note: RMSE difference between Benchmark and what we obtained can be partially attributed to the fact that the Benchmark models use 10-fold cross validation during training and testing whereas we use fixed partitions in order to enable apples-to-apples comparison of the two models' performance.

As shown in Table 3, CIGIN2 has a slightly better performance than DELFOS in terms of lower MAE, RMSE and MAPE. Therefore, when considering the results of these models for predicting the top solute - solvent pairs for a potential Covid-19 drug, CIGIN2 should be considered with greater weight.

While our models did not result in comparable RMSE in Benchmark models, the performance is reasonable and is partially due to the fact that we use less training data to train our models compared to cross-validation approach used in respective Benchmark models.

Discussion of Results

We will now discuss the models' predictions of the solvation free energies for the potential Covid-19 solutes - solvents pairs, namely the 138 solute x 28 solvent matrix.

Firstly, we consider which of the 28 solvents do best and the worst across the 138 solutes, i.e., have the lowest and highest solvation free energies predicted by the 2 models. Lower solvation energy means the solute - solvent pair is likely to be more effective as a composition of a Covid-19 drug. To obtain these results, an algorithm was implemented that calculates the weighted score of each solvent applied on the number of times its solvation energy was ranked 1st, or 2nd or 3rd for each solute. Weights applied were 0.5 for 1st rank, 0.3 for 2nd rank, and 0.2 for 3rd rank. The worst performing solvents were decided based on the number of times its solvation energy was ranked last for each solute. The best solvents yielding the lowest solvation energy by model are as follows:

Table 4a. Top 5 solvents and corresponding weighted scores for each model

Top Solvents (CIGIN2)	Weighted Scores (CIGIN2)	Top Solvents (DELFOS)	Weighted Scores (DELFOS)
Anisole	37.3	1-Pentanol	63.5

Isobutyl Acetate	37.0	1-Butanol	26.2
Butyl Acetate	29.4	Ethyl Formate	25.3
1-Pentanol	16.3	3-Methyl-1-butanol	12.7
3-Methyl-1-butanol	12.1	Water	6.1

According to our research, across all 138 solutes, the two models indicate some common best solvents namely, 1-Pentanol and 3-Methyl-1-butanol in top 5 solvents going by lowest free solvation energy. As noted earlier we will give more weightage to CIGIN2 model results because this model has lower error (e.g., MAPE, MAE, and RMSE) compared to DELFOS model.

Conversely, we also looked at the "worst" solvents i.e., ones with the highest solvation energy for each model.

Table 4b. Worst performing solvent(s) for each model

Model	**Worst Performing Solvent 1**	**Worst Performing Solvent 2**
CIGIN2	Water	-
DELFOS	Tert-Butylmethyl Ether	Water

According to our research, the worst solvents are Water as per CIGIN2 as well as DELFOS model, and additionally Tert-Butylmethyl Ether as per DELFOS model.

Next, we examined the best solvents for a set of solute types (e.g., anti-viral, anti-cancer, and anti-fungal) by model. The results are as follows:

Table 5. Top performing solvents corresponding to solute types

<table>
<tr><th rowspan="2">Model</th><th colspan="7">Top Performing Solvents</th></tr>
<tr><th colspan="4">Anti-cancer (n = 18)</th><th colspan="3">Anti-fungal (n = 5)</th></tr>
<tr><td>CIGIN2</td><td>Isobutyl Acetate</td><td colspan="2">Isopropyl Acetate</td><td>2-Butanol</td><td>Anisole</td><td>Isobutyl Acetate</td><td>Butyl Acetate</td></tr>
<tr><td>DELFOS</td><td colspan="2">Ethyl Formate</td><td colspan="2">Water</td><td>Ethyl Formate</td><td>1-Pentanol</td><td>Isobutyl Acetate</td></tr>
</table>

Table 5. (Continued)

Model	**Top Performing Solvents**

	Anti-malarial (n = 8)			Anti-psychotic (n = 5)		
CIGIN2	Isobutyl Acetate	3-Methyl-1-butanol	1-Pentanol	Isobutyl acetate	Anisole	Propyl acetate
DELFOS	1-Butanol	3-Methyl-1-butanol	1-Pentanol	Ethyl Formate	1-Pentanol	

Table 5. (Continued)

Model	Top Performing Solvents					
	Anti-viral (n = 9)			Anti-retroviral (n = 9)		
CIGIN2	Isobutyl Acetate	3-Methyl-1-butanol	1-Pentanol	Isobutyl acetate	3-Methyl-1-butanol	1-Pentanol
DELFOS	Water	Ethyl Formate	3-Methyl-1-butanol	1-Butanol	1-Pentanol	

According to our research, across 6 solute types, the two models indicate some common best solvents for Anti-fungal (Isobutyl Acetate), Anti-Malarial (3-Methyl-1-butanol, 1-Pentanol), Anti-viral (3-Methyl-1-butanol), and Anti-retroviral (1-Pentanol), but no common top solvents for Anti-Cancer and Anti-psychotic solute types. It is also interesting to find that best solvents differ across solute types. Going by lower MAPE, MAE, RMSE, we will recommend giving higher weightage to CIGIN2 model results.

Finally, we looked at solutes with highest pIC50 values, which indicates the most promising solutes to treat Covid-19. pIC50 value is a biological activity property of a solute (i.e., an API) which indicates the amount of API required to achieving 50% inhibition of the disease. It is one of the most important biological properties in drug design. We present here the list of top 3 best solvents for the top 5 solutes selected on the basis of highest pIC50 value.

The results of most biologically active solutes as measured by pIC50 value are presented in Table 6a for CIGIN2 and in Table 6b for DELFOS model.

Table 6a. CIGIN2 model - Top solvents corresponding to top solutes (according to pIC50 values)

Top pIC50 Solutes	Top Performing Solvents		
Amuvatinib (pIC50=7.7)	Isobutyl acetate	Anisole	Butyl acetate
Carboxyamidotriazole (pIC50=7.05)	Isobutyl acetate	Isopropyl acetate	3-Methyl-1-butanol
Ouabain (pIC50=7.01)	1-Pentanol	Butyl acetate	3-Methyl-1-butanol
Digoxin (pIC50=6.72)	Butyl acetate	Anisole	Isobutyl acetate

AI-10-49 (pIC50=6.72)	Anisole	Isopropyl acetate	Isobutyl acetate

Table 6b. DELFOS model - Top solvents corresponding to top solutes (according to pIC50 values)

Top pIC50 Solutes	**Top Performing Solvents**		
Amuvatinib (pIC50=7.7)	Isobutyl acetate	1-Pentanol	Ethyl formate
Carboxyamidotriazole (pIC50=7.05)	1-Pentanol	Ethyl formate	Water
Ouabain (pIC50=7.01)	1-Pentanol	1-Butanol	3-Methyl-1-butanol
Digoxin (pIC50=6.72)	1-Butanol	3-Methyl-1-butanol	1-Pentanol
AI-10-49 (pIC50=6.72)	Ethyl formate	1-Pentanol	1-Butanol

According to our research, across the top 5 solutes by pIC50 value, the two models indicate some common best solvents for Amuvatinib (Isobutyl acetate) and for Ouabain (1-Pentanol, 3-Methyl-1-butanol), but no common top solvents for Carboxyamidotriazole, Digoxin, and AI-10-49. Once again going by lower MAPE, MAE, RMSE, we will give higher weightage to CIGIN2 model results.

Advantages of our research approach

The in-silico approach as developed and discussed in this research has the advantage of speed. We are able to evaluate pairings of hundreds of solutes and dozens of solvents in a matter of hours, as opposed to weeks and months using traditional in-vitro methods. The two models we used were trained and tested in about a month, however post model-training, it took only hours for predicting solvation energies for the 138 solutes and 28 solvents. Given the scalability of the in-silico approach, we can easily expand the list of solutes and solvents without much time or cost impact.

Notwithstanding the initial cost of in-silico model development compared to in-vitro testing costs, once the AI models are trained and tested, future applications of the AI models cost much less time and money compared to traditional in-vitro (i.e., laboratory) approach.

Disadvantages of our research approach

There are two main limitations of the research work presented here. First, the models we used consider a pair of a solute and a solvent at a time. This works fine for evaluating long lists of solutes and solvents; however, it would be fruitful to incorporate more than one solvent, i.e., multiple excipients. In some cases, the interactions between multiple excipients can be important which need to be considered as well.

Second limitation of our approach presented here is that the solvation energy is also a function of the molar concentrations of the solute and solvent, i.e., amount of substance per unit volume of solution. We do not consider molar concentration in our present research.

Conclusions and Future Research Directions

This research is about discovering Covid-19 medicine from among known APIs with the most promising solubility. We focused on solvation energy as the key factor. We implemented two AI models, of which CIGIN2 model does somewhat better than DELFOS model in predicting solvation energy of solute-solvent pairs.

We are able to derive several conclusions regarding (a) best solvents across a list of solutes, (b) worst solvents across solutes, (c) best solvents by type of solute (e.g., Anti-viral, Anti-malarial, etc.), (d) best solvents for most biologically active solutes as measured by pIC50 value.

We also find that the two models yield significantly different results in terms of best solvents, although there are a few commonalities also. Overall given somewhat better performance of CIGIN2 model over DELFOS on test data, we recommend giving higher weightage to CIGIN2 model results.

The recommendations regarding top solvents from our research are best considered as a promising list for in-vitro study next. In other words, outputs from AI models in our research should be next experimentally tested in a laboratory (in-vitro work).

There are at least two promising areas for future research. One, to incorporate more than one solvent in the mix with a solute, and second, to incorporate molar concentrations of all solvents and the solute.

In conclusion, our research contributes to the use of AI models in drug discovery by quickly and efficiently discovering the most promising solvents for a set of solutes. We focused on Covid-19 APIs here as that problem is acute and in need of greater insight and help from the scientific community.

Our code for the models is available at https://github.com/Sampreeth04/COVID19-Drug-Design.

Acknowledgments

We thank the Department of Chemistry, Chemical Theory Center, and Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA for making available free-of-cost the Minnesota Solvation Database (MNSOL) version 2012 database.

References

1. Duarte Ramos Matos, G., Kyu, D. Y., Loeffler, H. H., Chodera, J. D., Shirts, M. R., & Mobley, D. L. (2017). Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data 2017, 62, 5*, 1559–1569. https://doi.org/10.1021/acs.jced.7b00104

2. Gilmer, J., Schoenholz, S., S Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). *Neural message passing for quantum chemistry*. 34th International Conference on Machine Learning-Volume 70 (pp. 1263–1272). JMLR.org.

3. Jaeger, S., Fulle, S., Turk, S. (2018). Mol2Vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling, 58, 27–35.* https://doi.org/10.1021/acs.jcim.7b00616

4. Kröger, L. C., Müller, S., Smirnova, I., & Leonhard, K. (2020). Prediction of Solvation Free Energies of Ionic Solutes in Neutral Solvents. *J. Phys. Chem. A 2020, 124, 20,* 4171–4181. https://doi.org/10.1021/acs.jpca.0c01606

5. Lim, H., and Jung, Y. (2019). Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci. 10:8306–8315.* https://pubs.rsc.org/en/content/articlelanding/2019/sc/c9sc02452b#!divAbstract

6. Marenich, A. V., Kelly, C. P., Thompson, J. D., Hawkins, G. D., Chambers, C. C., Giesen, D. J., Winget, P., Cramer, C. J., & Truhlar, D. G. (2012). Minnesota Solvation Database – version 2012, *University of Minnesota, Minneapolis*. https://doi.org/10.13020/3eks-j059

7. Morgan, H. L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation, 5*, 107–113. https://doi.org/10.1021/c160017a018

8. Pathak, Y., Laghuvarapu, S., Mehta, S., & Priyakumar, U. D. (2020). Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-Like Molecules. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 873-880. https://doi.org/10.1609/aaai.v34i01.5433

9. Pennington, J., Socher, R., Manning, C. (2014). *Glove: Global Vectors for Word Representation*. Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA (pp. 1532–1543).

10. Schuster, M., Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*, 2673–2681.

11. Vinyals, O., Bengio, S., & Kudlur, M. (2016). *Order matters: Sequence to sequence for sets*. ICLR 2016. https://arxiv.org/abs/1511.06391