A - PRESENTATION OF THE DOCTORAL PROJECT

PROJECT TITLE

PREDICTION OF MOLECULAR ASSEMBLY IN CRYSTALLINE FORMS OF ACTIVE PHARMACEUTICAL INGREDIENTS (APIS) USING AI MODELS – MACHINE LEARNING

PROJECT DESCRIPTION

Context

In recent years, Artificial Intelligence (AI) and Machine Learning have emerged as major tools for exploring chemical space and predicting complex molecular properties. By establishing nonlinear relationships between structure, electronic properties, and intermolecular interactions, these approaches open new perspectives for modeling phenomena that are difficult to address through purely experimental or theoretical methods, particularly in predicting the self-assembly and three-dimensional organization of molecules in the solid state.

Pharmaceutical active ingredients (Active Pharmaceutical Ingredients, APIs) are often in solid form as they are chemically stable, easy to administer and produce, and allow for precise dosing.[1] Medicines in solid form can exist either in crystalline or amorphous states, with a preference for the former due to the instability of many amorphous materials.[2] An API can be crystallized alone, as a single component, or with other molecules to form multi-component crystals, such as hydrates, solvates, salts, or co-crystals. For each of these phases, and for their various polymorphic forms that may exist, the different crystals of the same API will have different physicochemical properties, which can affect the stability and bioavailability of the drug.[3-5] Such diversity offers the possibility of controlling the physicochemical properties of the API while preserving the therapeutic activity of the active ingredient.

A co-crystal is a homogeneous crystalline solid formed by the association of two or more electrically neutral molecules (Fig. 1 left), which are present in precise stoichiometric amounts in the solid state and at room temperature.[6] The components of a co-crystal are held together in the crystalline phase by intermolecular (non-covalent) interactions such as hydrogen bonds, halogen bonds, chalcogen bonds, or π···π interactions.[7] It is possible to obtain hydrates or solvates of co-crystals if solvent molecules are incorporated into their crystal structures (Fig. 1 left).

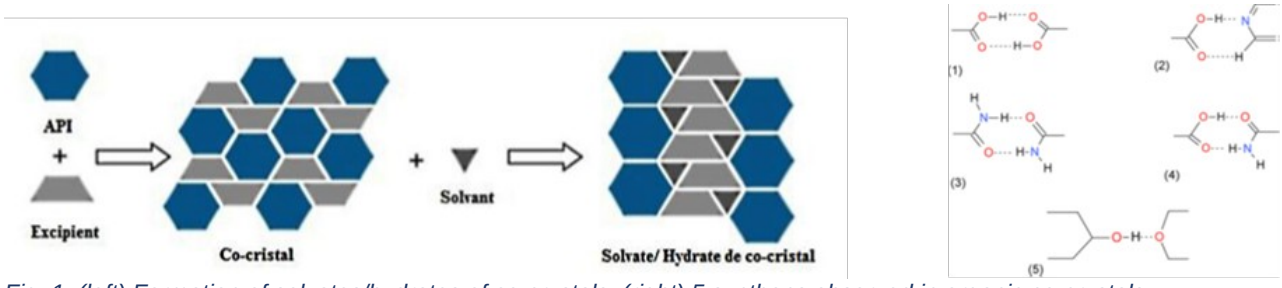

Fig. 1. (left) Formation of solvates/hydrates of co-crystals, (right) 5 synthons observed in organic co-crystals.

Co-crystals have a wide range of applications,[8] but the most interesting ones are in the pharmaceutical field.[3,9] The solubility of drugs is one of the most important biopharmaceutical properties, and its improvement is currently one of the main challenges of the pharmaceutical industry. In this regard, the formation of API co-crystals is one of the methods used to improve the bioavailability of drugs with APIs of low solubility.[10] Furthermore, in addition to improvements in solubility, dissolution rate, and bioavailability, the formation of co-crystals also improves other properties of APIs, such as stability, compressibility, flow, melting point, and hygroscopicity.[11] In pharmaceutical co-crystals, one of the components is an active pharmaceutical ingredient (API), and the other components are excipients, which facilitate drug administration but have no real therapeutic function. The molecules that can be used as excipients are listed in the GRAS (Generally Recognized As Safe) list published by the U.S. Food & Drug Administration (https://www.fda.gov/).

The hydrogen bond is the most commonly observed intermolecular interaction in the case of crystals of organic compounds containing acidic hydrogen atoms. The use of rules concerning the formation

of hydrogen bonds and molecular synthons, which are defined as recurrent intermolecular structural motifs in crystalline phases, can help in the design and analysis of co-crystals of organic compounds (Fig. 1 right). Studies on the competition between different supramolecular synthons have been conducted to facilitate the prediction of co-crystal formation.[12] The results of these studies suggest that the energetically most stable synthon is the one present in the arrangement of the formed co-crystal.[9] These hypotheses have been experimentally verified through comparisons of formed single-crystal structures.[13] The complementarity between molecular functional groups and their supramolecular synthons provides the prerequisites for the design of co-crystals. They facilitate the selection of appropriate molecules participating in the formation of co-crystals.

The crystalline solid of an API, whether formed by one or more components, can exhibit several forms associated with the phenomenon of polymorphism (Fig. 2). The different polymorphic forms have different physicochemical properties, as these depend on the solid-state structure, which varies according to the formed polymorph. Thus, by affecting the bioavailability of the active ingredient, the various polymorphic structures of an API manifest with (sometimes notable) differences in drug efficacy.[14] For a molecular crystal of an API, there can be a large number of polymorphs, and it is extremely difficult to know how many polymorphic forms a chemical compound can form. In the pharmaceutical field, polymorphism affects more than 80% of active ingredient molecules,[15] and about 30 to 50% of the best-selling pharmaceutical compounds on the market are polymorphic[16]. Therefore, the stakes of polymorphism are very high. It should be noted that, despite significant differences in intermolecular interactions, conformation, molecular position in space, and differences in the crystal structures of polymorphs of the same compound, their energy difference is generally very low, typically 2 to 16 kJ·mol⁻¹. Statistical studies have shown[17,18] that 85% of known polymorphs have an energy difference of less than 10% with an upper energy limit of 20 kJ·mol⁻¹, which underscores the difficulty in identifying them, especially since the crystallized forms are not necessarily those of lower energy or more stable.

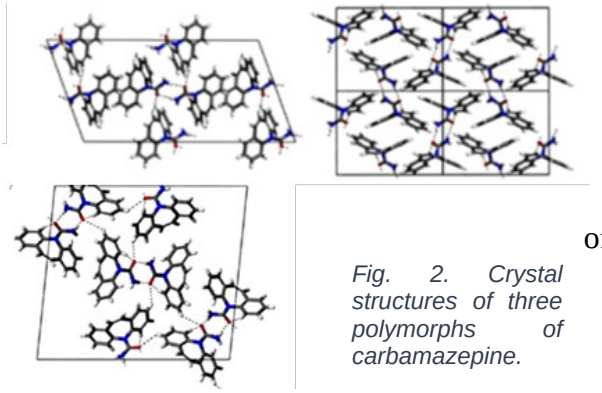

Fig. 2. Crystal structures of three polymorphs of carbamazepine.

Objectives

One of the most important aspects of the various crystalline forms of APIs and their polymorphs is that they can improve the properties of the active ingredient without changing its intrinsic biological activity. Molecular recognition, followed by self-assembly between the molecules forming the crystal, occurs via intermolecular interactions, often realized within synthons. Thus, the most commonly used excipients for the formation of pharmaceutical co-crystals have functional groups capable of building synthons with APIs. However, despite the use of synthons and phenomenological rules guiding the supramolecular synthesis of co-crystals, their formation cannot be predicted. Indeed, the problem is related to the 3D assembly required to form a crystal, and the use of complementary molecular groups capable of forming a synthon *only allows predicting part of the overall molecular assembly of the crystal*. Moreover, as indicated, *it is also very difficult to identify the existence of polymorphic forms from energy calculations*.

Recently, in collaboration with Dr. Sabeur Aridhi (LORIA, *Laboratoire Lorrain de Recherche en Informatique et ses Applications*), we have developed AI – Machine Learning models for predicting molecular properties (atomic charges and total molecular energy). To do this, we generated an SQL database with more than 455,000 entries from quantum DFT calculations of the molecular wave function (B3LYP/Def2TZVPP) obtained from more than 16,700 organic molecules from the *Cambridge Structural Database* (https://www.ccdc.cam.ac.uk/) containing H, C, N, O, F, P, S, Cl, Se, and Br atoms and their extended chemical types. After training, the model is capable of predicting atomic charges[19] with a standard deviation < 0.01*e* and the total energy of molecules with a standard deviation of 0.005 Hartree (13 kJ/mol), which should further improve

with a conformational exploration of the molecules. This research will result in a scientific publication.[20]

In this context, and using AI – *Machine Learning* methods, the objectives of this doctoral thesis project are:

- **Learn informative molecular representations** (descriptors / *embeddings*) from quantum data (atomic charges, densities, energies), allowing a fine description of intermolecular interactions.
- **Predict molecular properties** that describe intermolecular interactions.
- **Predict the 3D molecular assembly of an active ingredient (API)** (crystals, co-crystals, and polymorphs) using the previously predicted molecular descriptors.

Doctoral Project Timeline

The project involves developing a three-stage methodology: (i) prediction of atomic and molecular properties in the gas phase, (ii) prediction of molecular assemblies (orientations and intensities), particularly in one or more 3D layers around a central molecule, and (iii) from the assemblies obtained in (ii), determine the periodicity conditions on the obtained electronic descriptors to form a crystal.

This scientific project is highly innovative due to its methodology using AI-Machine Learning models for predicting molecular assemblies, particularly APIs. Therefore, skills in AI – Machine Learning are an important prerequisite for applying, while skills in quantum chemistry calculations and molecular modeling can be acquired during the thesis. Informatic developments and AI models for predicting molecular assemblies will be carried out in collaboration with Pr. Sabeur Aridhi (LORIA). Local (LORIA, CRM2) and regional (EXPLOR mesocenter) computing resources will be used.

References

[1] Hilfiker, R. (2006). Polymorphism: In the Pharmaceutical Industry. Wiley-VCH, Germany.
[2] Vippagunta, S. R., Brittain, H. G. & Grant, D. J. W. (2001). Adv. Drug Delivery Rev., 48, 3-26.
[3] Shan, N. & Zaworotko, M.J. (2008). Drug Discov. Today. 13, 440–446.
[4] Byrn, S. R., Pfeiffer, R., Stephenson, G. A., Grant, D. J. W. & Gleason, W. (1994). Chem. Mater. 6, 1148-1158.
[5] Byrn, S. R., Pfeiffer, R. R., Ganey, M., Hoiberg, C. & Poochikian, G. (1995). Pharmaceut. Res. 12, 945-954.
[6] Aakeroy, C. B. & Salmon, D. J. (2005). CrystEngComm.7, 439–448.
[7] Aakeroy, C. B. (1997). Acta Cryst. B53, 569–586.
[8] Steed, J. W. (2013). Trends Pharmacol. Sci. 34, 185–193.
[9] Blagden, N., Berry, D. J., Parkin, A., Javed, H., Ibrahim, A., Gavan, P. T., De Matos, L. L. & Seaton, C. C. (2008). New J. Chem. 32, 1659–1672
[10] Qiao, N., Li, M., Schlindwein, W., Malek, N., Davies, A., Trappitt, G. (2011). Int. J. Pharm. 419, 1–11.
[11] Lu, J. & Rohani, S. (2009). Org. Process Res. Dev. 13, 1269–1275.
[12] Etter, M. C. (1991). J. Phys. Chem. 95, 4601–4610.
[13] Vishweshwar, P., McMahon, J. A., Bis, J. A. & Zaworotko, M. J. (2006). J. Pharm. Sci. 95, 499–516.
[14] Brittain, H.G. & Grant, D.J.W. (1999). Effect of polymorphism and solid- state solvation on solubility and dissolution rate, in: H.G. Brittain (Ed.), Polymorphism in Pharmaceutical Solids. Marcel Dekker, New York.
[15] Giron. D. (1995). Thermochim. Acta. 248, l-59.
[16] Threlfall, T. L. (1995). Analyst. 120, 2435-2460.
[17] Gavezzotti, A. & Filippini. G. (1995). J. Am. Chem. Soc. 117, 12299-12305.
[18] Gavezzotti, A. (2007). Molecular Aggregation. Oxford University Press, New York.
[19] Bader, R. F. W. (1990). Atoms in Molecules − A Quantum Theory. Clarendon: Oxford, U.K.
[20] Autef, L., Aubert, E., Aridhi, S. & Espinosa, E. (2026). CrystEngComm (in preparation).

B – RESEARCH UNITS INVOLVED – HOST TEAMS

Research Units Involved :

- Laboratoire de Cristallographie, Résonance Magnétique et Modélisations (CRM2), UMR UL-CNRS 7036.
- Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), UMR UL-Lorraine INP 7503.

Host Teams:

- Structural Biology, Molecular Interaction Modeling, Crystal Engineering (BioMIMIC, CRM2)
- Artificial Intelligence and Big Data (LORIA – TELECOM Nancy)

C – MATERIAL, SCIENTIFIC, AND FINANCIAL CONDITIONS OF THE RESEARCH PROJECT

Both laboratories have the necessary computing resources (power and storage on our servers, as well as access to regional (MESOCENTRE) or national computing centers for heavier calculations) and infrastructure (office, computers, etc.) to welcome and support the person recruited for this research topic.

The research project is funded by the Université de Lorraine, reflecting its choice to support research through the awarding of doctoral contracts.

D – DOCTORAL PROJECT SUPERVISION

The interdisciplinary research work of this thesis will be carried out within the CRM2 and LORIA laboratories, under the supervision of:

- Pr. Enrique Espinosa (CRM2): enrique.espinosa@univ-lorraine.fr
- Dr. Emmanuel Aubert (CRM2): emmanuel.aubert@univ-lorraine.fr
- Pr. Sabeur Aridhi (LORIA): sabeur.aridhi@loria.fr

E – PROFILE AND SKILLS SOUGHT

Holder of a Master's degree (or equivalent) in AI, Data Science, Computer Science, Chemistry, or Physics, the candidate has strong skills in *Machine Learning* and scientific programming (Python). An interest in interdisciplinary approaches at the interface between AI and molecular modeling is essential. Knowledge of quantum chemistry or crystallography will be appreciated but is not mandatory.

Desired Profile:

✓ Master's in AI / Data Science / Computer Science / Chemistry / Physics

✓ Good level in *machine learning* / *deep learning* (PyTorch, scikit-learn, etc.)

✓ Interest in digital health and multimodal data

Environment: Applied research, collaboration with computer scientists – quantum chemists – crystallographers, real data, international publications.

F – APPLICATION

Send the following documents to the email addresses of the three supervisors (see section D above):

- CV
- Cover letter
- Transcripts from M1 and M2 (excluding internship grades for M1 and M2)