Optimisation de matériaux et dispositifs pour l'énergie à partir de concepts d'intelligence artificielle pour small data

K. KHOUSSA[1,2,3], Y.-A. CHAPUIS[1,2], N. LACHICHE[1,3]

[1]Laboratoire ICube, UMR CNRS et Université de Strasbourg

[2]MaCEPV (Équipe Matériaux et composants électroniques et photovoltaïque)

[3]SDC (Équipe Sciences des données et des connaissances)

Résumé

L'optimisation des procédés de fabrication des matériaux organiques pour l'énergie photovoltaïque est un processus qui manque de transparence sur la chimie et la physique qui agissent sur les matériaux et qui peuvent ou non conduire à une optimisation des performances du dispositif qui en dérive. Ceci est particulièrement pertinent pour les dispositifs photovoltaïques organiques (OPV) qui dépendent de relations très complexes entre les structures chimiques et les propriétés photovoltaïques. Aujourd'hui, l'intelligence artificielle (IA) peut saisir la complexité d'un dispositif et construire des modèles qui peuvent prédire l'efficacité de conversion. Des approches fondées sur les petits jeux de données ont été développées au cours de la dernière décennie. En particulier, de nouvelles approches telles que le design of experiments (DoE) combiné avec l'analyse de l'apprentissage artificiel permettent à l'expérimentateur d'utiliser les ressources rares plus efficacement, avec une probabilité plus élevée d'arriver à un véritable optimum. Dans ce travail, nous discuterons de l'utilisation de méthodes ML basées sur DoE pour l'optimisation des dispositifs OPV. Aussi, nous élargirons notre exploration des méthodes d'IA pour les petites données en utilisant des concepts d'apprentissage actif afin de surmonter les limites du DoE.

Mots-clés

Intelligence artificielle, apprentissage artificiel, plan d'expériences, cellules photovoltaïques organiques, petits jeux de données, apprentissage actif.

Abstract

The optimization of materials manufacturing processes for photovoltaic energy is a process that lacks complete transparency of chemistry and physics that relies on constraints imposed by the user that may or may not lead to an overall optimum. This is particularly relevant for organic photovoltaic devices (OPV) which depend on a very complex relationship between chemical structures and photovoltaic properties. Today, artificial intelligence (AI) can grasp the complexity of a device and build models that can effectively predict and achieve optimal conversion efficiency. Small data approaches have grown significantly over the past decade. Recently, new approaches such as the Design of experiments (DoE) combined with machine-learning analysis allow the experimentalist to use scarce resources more efficiently, with a higher probability of achieving a true optimum. In this work, we will discuss the use of DoE-based ML methods for optimizing OPV devices. In addition, we will expand our exploration of AI methods for small data using active learning concepts to overcome DoE limitations.

Keywords

Artificial intelligence, machine learning, design of experiments (DoE), organic photovoltaic (OPV) cells, small data, active learning.

1. Introduction

L'intelligence artificielle (IA) est aujourd'hui présente dans de nombreux domaines des sciences, tels que la physique, la chimie, la biologie, la santé, et bien d'autres. On l'emploie également de plus en plus dans tous les secteurs de l'industrie, tels que l'automobile, la construction aéronautique, les semi-conducteurs, l'énergie et bien d'autres encore [1]. Cet intérêt pour l'IA s'explique en grand partie par les progrès dans le traitement des données massives, aussi appelé *big data*, lesquelles permettent une nouvelle approche pour analyser les processus industriels mais pose également le problème de la redondance de l'information, c'est-à-dire de l'information qui n'est pas strictement nécessaire pour définir les modèles de prédiction en toute confiance [1]-[2]. Il existe une alternative au *big data* par la collecte d'ensembles de données plus petits mais plus riches et plus sures en information. Cette approche, plus connue sous l'appellation *small data*, représente un vecteur d'innovation récent et prometteur mais encore complexe à traiter en IA [3]. Par exemple, la production de semi-conducteurs nécessite d'énormes quantités de données afin de pouvoir contrôler, améliorer et gérer la complexité des procédés de fabrication, d'autant plus si l'on souhaite accompagner les faibles latences des marchés. Ainsi, le concept de *small data* s'avère comme une solution d'avenir, bien que ce dernier reste une voie de recherche complexe de l'IA, et encore peu diffusée dans la littérature [4]. Récemment, des recherches en *small data* ont montré que la combinaison d'un simple plan d'expériences (*design of experiments - DoE*) avec un concept classique d'apprentissage artificiel (AA) (*machine learning - ML*) permettait d'accélérer l'optimisation de dispositifs à base de matériaux complexes pour l'énergie [2]. Cette méthode prometteuse reste encore à explorer, sachant que les performances prédictives des modèles d'apprentissage restent encore très en dessous des exigences industrielles [5]–[7]

Dans cet article, nous proposons un état de l'art sur les concepts de *small data* employés dans l'optimisation des matériaux et dispositif dérivés. Nous nous intéresserons

particulièrement au domaine des matériaux organiques pour l'énergie photovoltaïque (*organic photovoltaic - OPV*). En effet, l'organique présente la particularité de produire des cellules photovoltaïques par voie chimique ou liquide, ce qui le rend très attractif en termes d'intégration et de coût de production. Par contre, la complexité chimique de ces matériaux impose encore aujourd'hui un traitement long, délicat et nécessitant de nombreuses variables. Dans ce contexte, l'IA peut s'avérer comme une solution pour accélérer les développements dans cette technologie, laquelle se base sur des jeux de données réduits propres au *small data*.

2. Contexte scientifique

2.1. Matériaux : *OPV*

Tout comme leurs parents inorganiques, les cellules OPV utilisent l'effet photovoltaïque pour transformer l'énergie lumineuse en électricité. Par contre, les matériaux organiques ont l'avantage de produire des dispositifs beaucoup plus légers, flexibles, transparents, respectueux de l'environnement, et s'intégrant avec des substrats très variés. Ils sont également moins chers à produire grâce aux technologies de fabrication développées dans l'industrie de l'électronique flexible (e.g. OLED, etc.) [4]. Par contre, l'OPV reste encore aujourd'hui peu bénéfique en termes de rendement énergétique, durée de vie, stabilité, dégradation par exposition à l'oxygène, et dépendance à l'énergie solaire [5], [8], [9]. Pourtant, récemment, grâce à la découverte de nouveaux matériaux, l'OPV semble avoir rompu de nombreuses barrières technologiques telles que le rendement énergétique et de stabilité, ce qui les placent aujourd'hui en compétition directe avec les solutions PV classiques à base de matériaux inorganiques [5], [9].

Technologiquement, les matériaux OPV sont constitués de matériaux donneurs d'électrons et accepteurs d'électrons plutôt que de jonctions classiques *pn* semi-conductrices. Les cellules OPV sont ainsi fabriquées à partir d'une couche active contenant un polymère actif donneur d'électron et un accepteur d'électrons à base de molécules fullerènes ou non fullerènes. On parle de cellules solaires en polymère à hétérojonction en vrac ou BHJ (bulk heterojunction) [2]. La figure 1 présente la structure par couches d'une cellule OPV à BHJ.

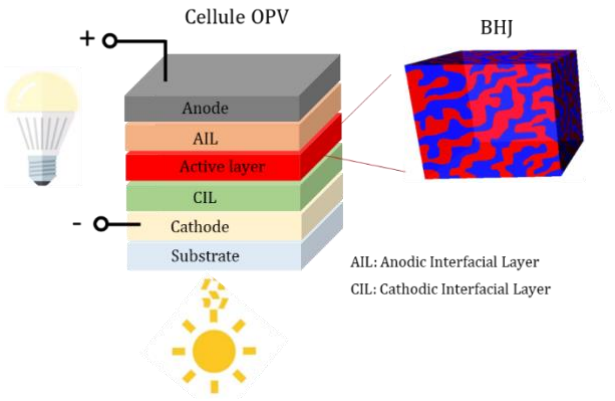

Figure 1 : Structure d'une cellule OPV à BHJ pour la conversion électrique.

Les BHJ nécessitent un contrôle sensible de la morphologie des matériaux à l'échelle nanométrique. Les variables importantes comprennent les matériaux, les solvants et le rapport pondéral donneur-accepteur, ce qui constitue toute la complexité chimique de la couche active photovoltaïque. De plus, la cellule OPV doit aussi prendre en compte la conception et l'adaptation des couches d'interface et des électrodes pour compléter le dispositif photovoltaïque [9]–[11]. Chacune des couches supplémentaires de la cellule OPV va entraîner de nouvelles variables expérimentales qui vont d'autant plus contraindre le travail d'optimisation de ces dispositifs énergétiques.

Récemment, de nouveaux matériaux ont été développés à partir d'accepteur non fullerène (*NFA*), ce qui a permis une progression spectaculaire des performances des dispositifs OPV. Dans la littérature, il a déjà été enregistré des rendements énergétiques supérieurs à 18% pour des durées de vie d'au moins 10 ans [9]. Ces résultats s'approchent et même dépassent les performances des cellules classiques PVs à couche minces à base de silicium. Cependant, pour poursuivre ces progrès, l'OPV doit à présent élaborer et optimiser une variété de matériaux type NFA, ce qui nécessitera de lourds investissements de recherche et de développement, aussi bien académiques qu'industriels. C'est dans ce contexte que l'IA et le *small data* offrent au secteur photovoltaïque une opportunité d'accélérer la découverte et d'optimiser les matériaux et dispositifs OPV [2].

2.2. Approche expérimentale : *DoE*

En matériaux, la plupart des découvertes ont été obtenues de manière empirique, généralement par le biais d'une approche expérimentation de type « *Edisonienne* » ou dite « à une variable à la fois » (*One factor at a time - OFAT*) [2]. Pourtant, cette méthode a de nombreux défauts car les caractéristiques des systèmes basés sur les matériaux ne sont ni simples, ni non corrélées, ce qui entraîne une quantité exhaustive du nombre de variables à expérimenter. En effet, il faut savoir que la modification d'une variable expérimentale peut avoir de multiples effets imprévus en raison de l'inter-connectivité entre leurs propriétés des matériaux. Généralement, on utilise une approche expérimentale basée sur les principes de DoE, lesquels englobent de nombreux calculs statistiques et permettent une analyse multi-variables. On peut donc tester et optimiser plusieurs variables simultanément, accélérant ainsi le processus de découverte et d'optimisation tout en économisant du temps et de précieuses ressources. Dans ce domaine, les travaux de Fisher [12], Box et Wilson [13] font depuis longtemps références dans le domaine de l'approche par *DoE*. Plus récemment, on mentionnera la méthode de Taguchi [14].

Bien que la méthode par DoE ait déjà fait ses preuves, cette dernière reste fondamentalement dépendante de l'expérimentateur, malgré le semblant d'automatisation. En effet, si le choix des DoEs se base sur des règles strictes d'échantillonnage, qui permettent d'explorer un grand nombre de paramètres dans un espace multidimensionnel, la méthode utilise un volume de données où l'expérimentateur est tenu de sélectionner les facteurs d'entrée qui doivent être inclus dans les expériences [12], [15]. Il est donc nécessaire d'avoir une bonne connaissance du processus d'élaboration du matériau ou du dispositif dérivé.

2.3. IA : Concept de *small data*

Il existe un intérêt croissant de l'emploi des concepts d'IA dans la recherche des matériaux et des dispositifs dérivés. Cependant, pour des raisons de coûts matériel et de développement, les jeux de données produits dans l'étude des matériaux sont généralement plus petits et plus diversifiés comparativement à d'autres domaines[3], [5], [16]. Si la réduction de la taille des jeux de données peut sembler être un désavantage dans l'élaboration des modèles d'IA, il faut aussi considérer les avantages, comme la fiabilité, l'accessibilité, la compréhension et l'exploitation des données [8]. Par définition, les modèles d'IA appliqués aux matériaux font souvent références au concept de *small data*. Dans l'approche de l'IA pour les matériaux, si le concept de *small data* reste encore peu développé et présent dans la littérature, il commence tout de même à bénéficier des développements dans le domaine général de l'IA , et l'on voit de plus en plus d'articles faisant mention de ce thème [3], [16].

Récemment, une étude publiée par Zhanh et al. [3] présente des travaux permettant d'analyser de manière exhaustive l'interaction entre la disponibilité des données sur les matériaux et la capacité prédictive des modèles d'AA. Dans un premier temps, les auteurs donnent une représentation des erreurs de précision prédictive en fonction de la taille des jeux de données qu'ils ont collectées dans la littérature, comme le reprend la figure 2.

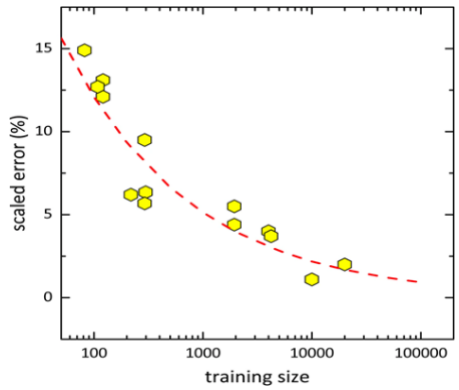

Figure 2 : Représentation des erreurs de précision prédictive en fonction de la taille des jeux de données, comme collectées dans la littérature [3].

Par la suite, une stratégie pour améliorer la précision prédictive des modèles classiques d'AA appliquée à des petits jeux de données pour études de matériaux est proposée. Les auteurs analysent l'interaction fondamentale entre la disponibilité des données sur les matériaux et la capacité prédictive des modèles d'AA. Leurs travaux révèlent que l'augmentation de la précision se fait au prix d'un degré de liberté (*degree of freedom - DoF*) du modèle. Pour remédier à ce problème, les auteurs proposent d'incorporer dans l'espace des caractéristiques un coefficient appelé « estimation brute de la propriété » (*crude estimation of property - CEP*). Puis, ils présentent trois cas d'étude de prédiction sur des matériaux : (i) bandes interdites de semiconducteurs; (ii) conductivité thermique; (iii) module d'élasticité des zéolithes. A travers ces exemples, les auteurs valident leur stratégie en démontrant que le CEP a effectivement amélioré la précision prédictive des modèles d'AA[2], [4].

3. Approche par apprentissage artificiel

3.1. Combinaison DoE-AA

Récemment, l'équipe du professeur J. Buriak de l'Université d'Alberta (Canada) a proposé de combiner les données de plan d'expériences (DoE) avec un modèle d'AA pour optimiser les performances de rendement d'un matériau pour OPV [2]. Un premier DoE est choisi pour initier le premier modèle d'AA à partir d'un jeu de seulement 16 conditions expérimentales. Puis, après visualisation des prédictions de la cible par cartographie (*mapping*), le DoE est affiné et le modèle peut être optimisé autant de cycles que nécessaire (boucle d'affinement). Le flot de la méthode appelée DoE-AA est illustrée sur la figure 3.

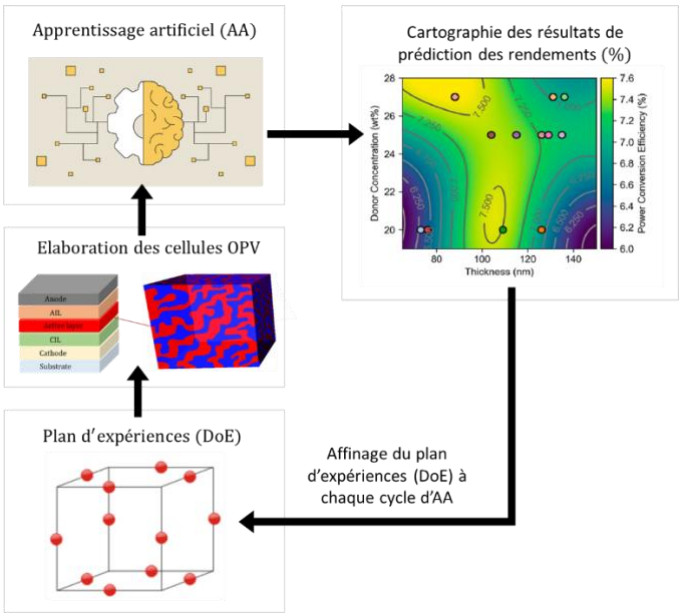

Figure 3 : Flot de la méthode DoE-AA pour l'optimisation des performances de rendement des matériaux pour cellules OPV

La méthode DoE-AA ne permet pas d'automatiser entièrement le processus d'IA mais il réduit significativement la dépendance à la décision de l'expérimentateur. En effet, si le choix du premier plan d'expériences (DoE) reste à l'initiative de l'expérimentateur, les autres cycles de décision sont fortement assistés par l'aide cartographique, laquelle ne nécessite pas intrinsèquement de connaissance en matériaux OPV. A noter que l'intérêt de cette méthode ne se limite pas aux matériaux OPV puisque plusieurs travaux de recherche s'en sont depuis inspirées pour diverses applications, telles que les électrodes transparentes [1]–[5], [8]–[11], [17], [18].

3.2. Etude des biais en AA et solutions

La méthode DoE-AA décrite dans la section précédente ouvre la voie de l'automatisation des processus d'optimisation des matériaux à partir de petits jeux de données. Cependant, la dépendance à l'utilisateur reste un point critique de la méthode. Dans l'article de review de Zhao et al. [2] les auteurs analysent plusieurs études combinant petites tailles de données

expérimentales et modèles d'AA pour l'optimisation de matériaux OPV. Ces études montrent que, généralement, les ensembles de données expérimentaux sont quelque peu biaisé, cela créant des difficultés pour évaluer objectivement les performances des modèles d'AA. Si les biais en IA sont un problème bien réel, c'est encore plus vrai dans l'élaboration et la caractérisation des matériaux, comme le montre notre étude en OPV.

Finalement, pour remédier aux biais et améliorer les modèles d'AA, les auteurs proposent un cycle de travail en quatre étapes combinant découverte, optimisation expérimentale et concepts d'AA: (1) Génération des données expérimentales en automatisant la procédure; (2) Sélection de descripteurs; (3) Analyse par AA; (4) Découverte des conditions expérimentales. Ce flot de travail est présenté sur la figure 4.

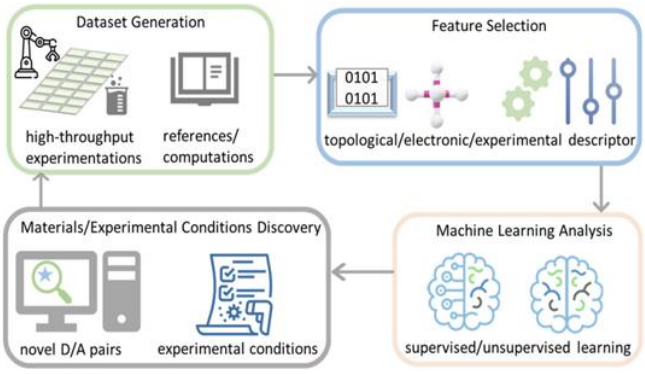

Figure 4 : Cycle de travail en quatre étapes combinant découverte, optimisation expérimentale et concepts d'AA [5]

Le cycle de travail proposé par Zhao et al. [5] se base sur l'élaboration d'un ensemble de données valides comme une condition préalable à toute méthodologie impliquant les concepts d'AA. Ces données sont liées à l'évaluation de la synthétisabilité des matériaux OPV et au choix des protocoles de fabrication de ces matériaux, ce qui est très coûteux en termes de matériel et temps de développement. Pour remédier à cette problématique de coût, le flot propose de combiner l'automatisation de la synthèse des matériaux avec un criblage virtuel assisté par AA avec une caractérisation à haut débit, et ceci en engageant de faibles tailles de données (de 70 à ~1800 données) [10]. On retiendra également la notion de laboratoire « autonome », ou « plate-formes robotisées » introduit pour réduire la dépendance du processus à la reproductivité des expériences. Finalement, ce cycle ouvre une voie de développement très prometteuse.

Un exemple d'application de cette approche par étape expérimentale automatisée est rapporté dans l'article de Du et al. [10]. Dans ce travail, les auteurs prennent en compte jusqu'à 10 variables expérimentales dans l'élaboration de cellules OPV (e.g. ratio donneur:accepteur, concentration, vitesse de dépôt, additif, solvant, température et temps de recuit, interfaces, procédés des interfaces) pour produire et caractériser automatiquement en moins de 24 heures plus de 100 cellules OPV à base de matériaux NFA de type PM6:Y6 (matériau référence pour l'OPV depuis 2019). La figure 5 présente la plate-forme auto-développée appelée AMANDA Line One pour la fabrication et la caractérisation automatique à haut débit de cellules OPV NFA. Les données expérimentales sont ensuite traitées par un modèle d'AA, lequel va permettre a d'identifier les critères les plus impactant dans l'optimisation des performances de rendement des cellules OPV.

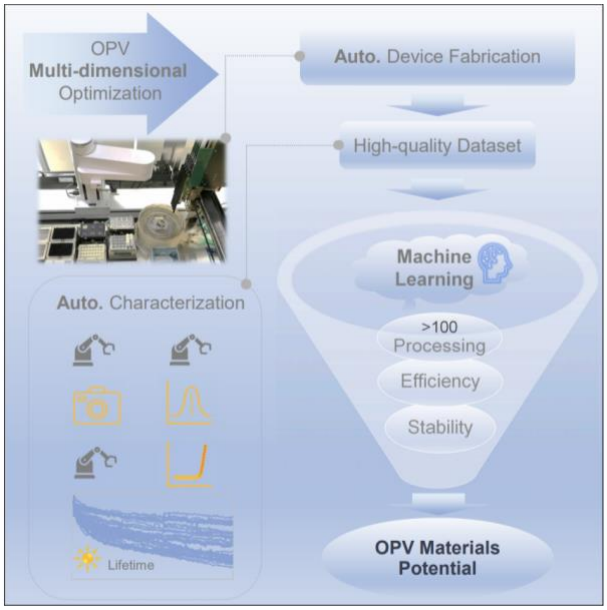

Figure 5 : Plate-forme auto-développée appelée AMANDA Line One pour la fabrication et la caractérisation automatique à haut débit de cellules OPV NFA [10].

3.3. Modèles d'AA pour small data

Dans la littérature, il n'existe pas, à notre connaissance, d'articles traitant de la pertinence de certains modèles d'AA pour le traitement de jeux de données de petites tailles. Par contre, lorsque nous manquons de donnés, il est courant que les modèles d'AA entraînent des problèmes, tels que :

- la précision de prédiction (*prediction accuracy*) ;
- la vitesse d'apprentissage (*training speed*) ;
- le surapprentissage (*overfitting*) :
- l'incapacité à valider le modèle (*inability to valide the model*)

Par conséquent, le choix de l'algorithme d'AA devra être toujours chercher à répondre positivement à ces critères.

Dans cette revue scientifique, nous avons relevé que pour gérer la faible quantité de données expérimentales dans le domaine des matériaux OPV, la plupart des auteurs développait des stratégies accompagnant les modèles d'AA, comme la combinaison de modèle d'AAA avec un plan d'expérience (DoE) ou un critère d'évaluation (CEP). Un autre point que nous avons relevé est l'emploi d'algorithmes d'AA avancés, tels que le Support vector machine (SVM) et le Gaussian process regressor (GPR).

Dans les sections suivantes, nous décrirons brièvement les algorithmes d'AA SVR et GPR, en mettant en avant leur adéquation avec les jeux de donnés réduits.

3.3.1. Support Vector Regression (SVR)

En général, le Support vector machine (SVM) est un

algorithme d'apprentissage supervisé linéaire qui est largement utilisé dans diverses industries, médecine, énergie, etc. pour résoudre les tâches de régression et de classification [16]. Cependant, en raison de la possibilité d'utiliser différents noyaux (*kernels*), cette méthode est également utilisée pour résoudre des tâches non linéaires, ce qui . Dans le cas de données quantitatives, comme employées dans l'étude de matériaux OPV, on parlera de l'algorithme Support vector regression (SVR). Les principaux avantages de l'algorithme SVR sont énuméré ci-dessous :

- fournir un travail efficace avec les petits jeux de données (*small data*) ;
- afficher de bons résultats lorsqu'il travaille dans un espace de dimensions supérieures ;
- produire une résolution sans ambiguïté.

Une série de fonctions mathématiques sont utilisées dans SVR pour convertir les données qu'il reçoit en entrée, appelé *Kernel*, dans le formulaire requis. Des prédictions non linéaires peuvent être faites dans le modèle créé par ces fonctions du kernel. Dans la littérature nous pouvons trouver pas mal de type de kernel mais dans notre cas avec les petits jeux de données, nous pouvons mentionner : les kernels polynomiales et FBR et linéaire. La relation (1) est utilisée pour le processus de normalisation:

$$y = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

où : y est la valeur normalisée de x_i, x_{max} est la valeur maximale de x_i, et x_{min} est la valeur maximale de x_i.

Les relations de (2) à (4) sont utilisées pour le calcul des fonctions des kernels polynomial, RBF et linéaire :

Kernel polynomial :

$$k(x_i, x_j) = (x_i . x_j + 1)^d \qquad (2)$$

Kernel RBF :

$$k(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad (3)$$

$$\text{où : } \gamma = \frac{1}{2\sigma^2} \quad \text{pour } \gamma > 0$$

Kernel linéaire :

$$k(x_i, x_j) = x_i^T x_j \quad (4)$$

Bien que l'algorithme SVR fonctionne efficacement avec les petits jeux de données, cette méthode n'est pas toujours très précise. La taille de l'ensemble de données, le bruit ou les valeurs aberrantes jouent également un rôle important pour son rendement [16].

3.3.2. Gaussian process regressor (GPR)

L'algorithme Gaussian process regressor (GPR) est basé sur la théorie des probabilités bayésienne et a des liens très étroits avec d'autres techniques de régression, comme la régression des crêtes du kernel (*KRR*) et la régression linéaire avec des fonctions radiales [19]. Les modèles de régression basés sur les processus gaussiens sont simples à mettre en œuvre, flexibles, entièrement probabilistes, et donc un outil puissant dans de nombreux domaines d'application [20]. Un processus gaussien est une sélection (peut-être infinie) de variables aléatoires pour lesquelles tout sous-ensemble fini de ces variables a une distribution gaussienne conjointe [21], [22]. Les variables sont généralement indexées par l'ensemble x, donc ensemble les variables $f(x)$ sont considérées comme une fonction (stochastique) sur l'ensemble d'index. Pour tout sous-ensemble fini d'indices x_1, x_2, ..., x_n, nous avons :

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \sim$$

$$\mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \dots & k(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \dots & k(x_n,x_n) \end{bmatrix}\right) \qquad (5)$$

avec $\mu(x)$ est la fonction moyenne, et $k(x_i, x_j)$ est le kernel. Le kernel comme nous avons dit est une fonction semi-définie positive qui spécifie la matrice de covariance pour tout sous-ensemble fini de variables [23].

3.4. Performance de prédiction

Les mesures de rendement du modèle, comme le r, le coefficient de détermination (R^2), l'erreur racine moyenne carré (RMSE) , erreur quadratique moyenne (MSE), l'erreur moyenne absolue en pourcentage (MAPE) et l'erreur moyenne absolue (MAE) peuvent être utilisé pour les modèles de régression, ce qu'on peut les définir comme suit [5] :

$$r = \frac{\sum_{i=1}^{N}(R_i - \bar{R}_\iota) \times (P_i - \bar{P}_\iota)}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R}_\iota)^2 \times \sum_{i=1}^{N}(P_i - \bar{P}_\iota)^2}} \qquad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(R_i - \bar{P}_\iota)^2}{N}} \qquad (7)$$

$$\text{MSE} = \frac{\sum_{i=1}^{N}(R_i - P_i)^2}{N} \quad (8)$$

$$\text{R}^2 = 1 - \frac{\text{MSE}}{\text{var}(R_i)} \quad (9)$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i - R_i|}{|R_i|} \quad (10)$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}|R_i - P_i|}{N} \quad (11)$$

avec :

- N : nombre de points de données dans l'ensemble de données.
- R_i et P_i : valeur réelle et prévue.
- $\bar{R}_\iota$ et $\bar{P}_\iota$: valeurs moyennes pour la valeur réelle et prévue, respectivement.
- $\text{var}(R_i)$: variance des données de l'échantillon.

4. Approche par apprentissage actif

Dans cet article, nous proposons d'étudier le concept d'apprentissage actif (*active learning - AL*), lequel peut présenter de nombreux avantages au traitement des petits jeux de données [24].

4.1. Notions d'AL

L'apprentissage actif ou AL, aussi connu sous le nom de conception expérimentale optimale ou apprentissage par requête, est un sous-domaine de l'AA et plus généralement de l'IA. Le concept d'AL a été introduit en 1988 par D. Angluin [25], et a initié de nombreuses études jusqu'à une synthèse plus complète par B. Settles [26] en 2009.

Le concept clé de l'AL est que si l'algorithme d'apprentissage est autorisé à choisir les données à partir desquelles il va apprendre, il sera plus performant avec moins de données annotées. Les systèmes d'AL tentent d'éliminer le manque de données « étiquetées » (ou « annotées ») en demandant des « requêtes » sous forme d'« instance non étiquetées » qui doivent être « annotées » par un oracle. Le figure 6 illustre le concept d'AL via l'ensemble des éléments de traitement :

- jeu de données « étiquetées » (*labeled training set* $\mathcal{L}$) ;
- algorithme d'AA (*machine learning model*) ;
- stratégie de requête (*select queries*) ;
- jeu de données « non-étiquetées » ou « pool » (*unlabeled pool* $\mathcal{U}$) ;
- oracle.

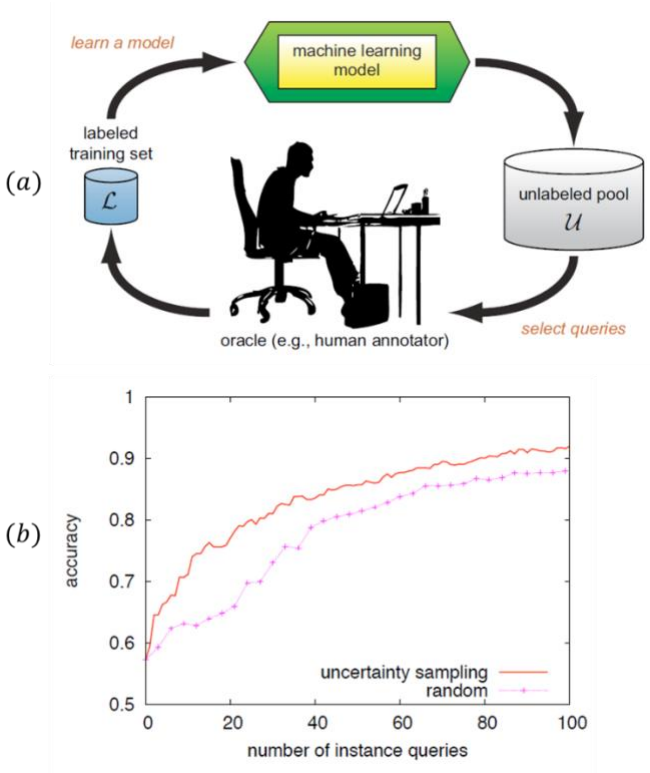

Figure 6 : (a) Flot de fonctionnement du système d'AL selon B. Settles. (b) Exemple de courbes d'apprentissage par AL montrant l'amélioration de la précision de prédiction d'un modèle quelconque en fonction du nombre de requêtes d'instance [26].

Il existe plusieurs scénarios dans lesquels l'algorithme peut interroger l'oracle, les trois principaux scénarios sont :

- la synthèse de requêtes par adhésion ;
- l'échantillonnage sélectif basé sur le flux ;
- l'AL basé sur un large jeu de données dites « non étiquetées » (*pool*).

Les systèmes d'AL doivent échantillonner des instances à partir d'un ensemble d'instances « non étiquetées » et utiliser une stratégie de requête pour décider s'il faut interroger un oracle (annotateur humain ou un appareil qui renvoie la véritable étiquette de l'instance) pour obtenir l'étiquette de l'instance ou pour l'abandonner.

En résumé, l'AL consiste à filtrer, au long d'itérations successives, les données les plus pertinentes à faire « étiqueter » ou annoter par un opérateur humain. Les nouvelles données annotées sont ajoutées aux précédentes pour construire un nouveau modèle qui sera lui-même utilisé pour affiner la sélection de nouvelles données à étiqueter. La figure 7 reprend cette démarche en formant un cycle d'apprentissage actif ou boucle d'Active Learning.

Comme le montre les exemples de la section précédente, l'approche par AL va permettre de réduire l'impact de l'étiquetage, et par conséquent de minimiser le nombre d'expériences pour obtenir le modèle d'apprentissage le plus précis.

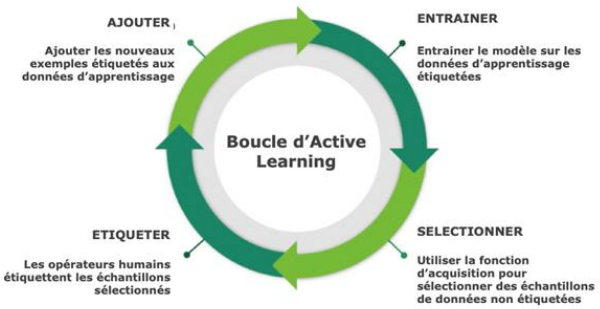

Figure 7 : Représentation de la boucle d'Active Learning [27]

4.2. Exemple d'application

Pour illustrer l'application de l'approche d'AL, nous présenterons les travaux de Lookman et al. [7], lesquels montrent l'utilisation des modèles et stratégies d'AL pour accélérer la découverte de nouveaux matériaux électroformables pour dispositifs piézoélectriques. Cette étude se base sur l'utilisation de données « non étiquetées », ou *pool*, correspondantes à près de 605 000 compositions d'un matériau électroformable qui pourra être appliqué au piézoélectrique. A chaque itération, 4 compositions du matériau seront extraites des données contenues dans le *pool*, *via* différentes stratégies de requête (exploitation, exploration, compromis entre les deux premiers et sélection aléatoire). Puis, un dispositif piézoélectrique sera élaboré et caractérisé pour chaque composition du matériau électroformable. Les résultats seront appelés « données d'itération », lesquelles seront ensuite « étiquetées » et intégrées au modèle d'AA. une nouvelle itération sera exécutée.

La boucle d'Active Learning et les résultats d'optimisation sont présentés sur la figure 8.

5. Conclusion

Cet article aborde le domaine de l'IA appliquée à l'optimisation de matériaux et dispositifs pour l'énergie, où la complexité et le coût élevé des processus restreignent le nombre d'expériences et de données disponibles. A travers une sélection dans la littérature récente, le thème combinant IA et petits jeux de données (*small data*) a été exploré dans le cas de matériaux pour le photovoltaïque organique (OPV) et de dispositif piézoélectrique. Deux approches d'IA ont été étudiées : (i) l'approche par AA ; (ii) l'approche par AL.

La première approche a permis de mettre en évidence l'amélioration des prédictions d'optimisation énergétique de matériaux OPV, en combinant algorithmes d'AA et méthodes classiques par plan d'expériences (DoE). On pourra retenir les travaux du professeur J. Buriak (2018), qui ont prouvé l'intérêt du concept, lequel fait aujourd'hui références. On notera que la tendance actuelle est le renforcement de cette approche par l'automatisation des procédés d'élaboration des matériaux OPV (Du et al., 2021), (Zhao et al., 2022).

La seconde approche, initiée par Yuan et al. (2018), propose le traitement de petits jeux de données par boucle d'AL dans le cas de cherche sur des matériaux pour dispositifs piézoélectriques. Ce concept semble très prometteur dans le domaine du *small data* car il permet de limiter le nombre d'expériences et d'améliorer les précisions de prédiction.

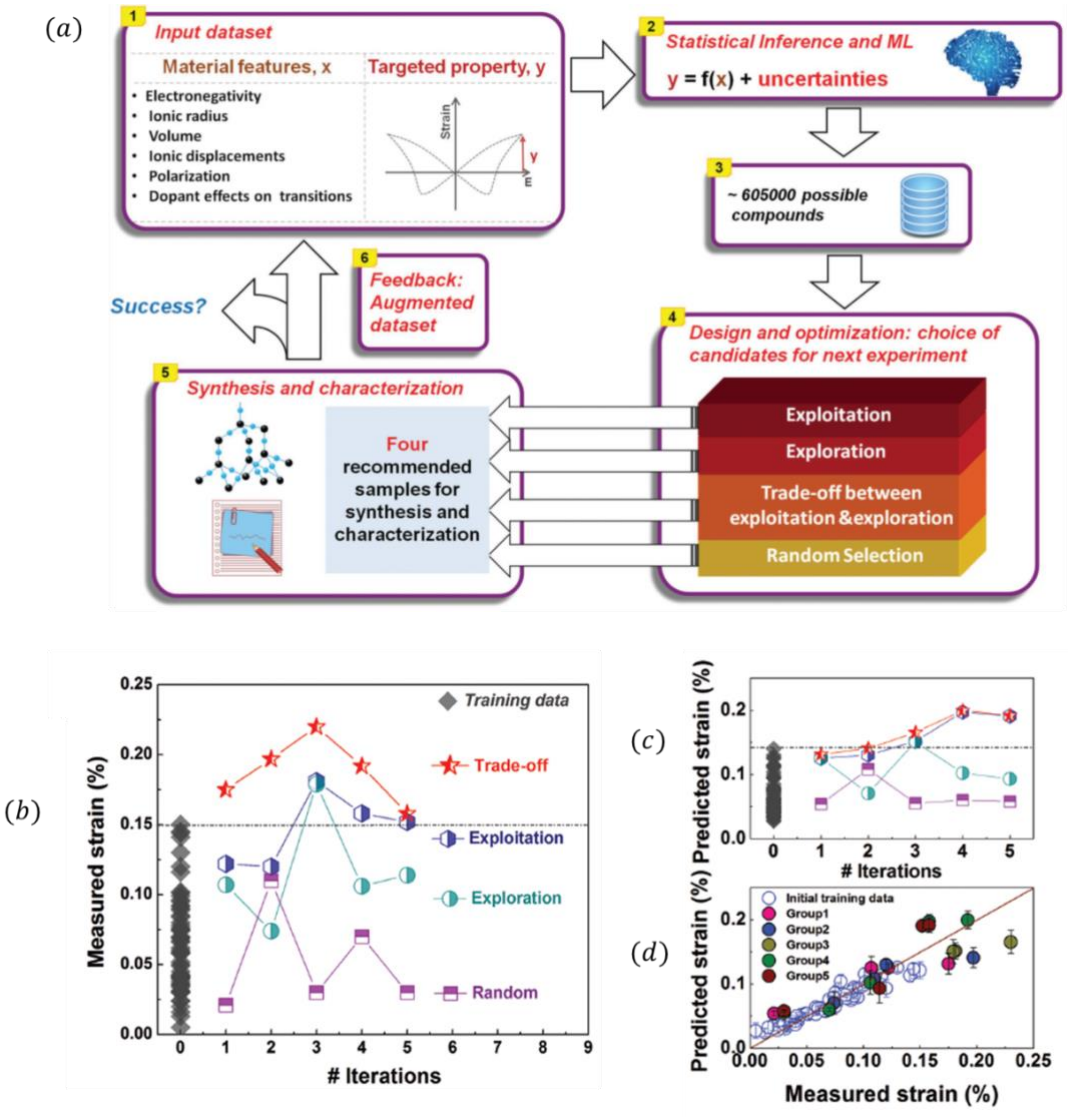

Figure 8 : (a) Boucle d'AL pour une découverte accélérée basée sur l'AA et une conception expérimentale optimale pour guider les expériences de manière itérative dans la recherche de piézoélectriques performants à grandes électrodéformations. Les composés sont synthétisés en suivant et comparant les prédictions de 4 stratégies (exploitation, exploration, compromis entre les deux premiers et sélection aléatoire). (b) Comparaison expérimentale des 4 méthodologies de conception montrant que le compromis entre l'exploration et l'exploitation fonctionne mieux à chaque itération que les autres stratégies pour trouver le composé avec les plus grandes électrodéformations. (c) Prédictions. (d) Les électrodéformations prédites et mesurées des nouveaux composés synthétisés sont en accord raisonnable et donner confiance dans la qualité du modèle d'inférence. [26]

Bibliographies

[1] L. Wei, X. Xu, Gurudayal, J. Bullock, and J. W. Ager, “Machine Learning Optimization of p-Type Transparent Conducting Films,” *Chemistry of Materials*, vol. 31, no. 18, pp. 7340–7350, Sep. 2019, doi: 10.1021/acs.chemmater.9b01953.

[2] B. Cao *et al.*, “How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics,” *ACS Nano*, vol. 12, no. 8. American Chemical Society, pp. 7434–7444, Aug. 28, 2018. doi: 10.1021/acsnano.8b04726.

[3] Y. Zhang and C. Ling, “A strategy to apply machine learning to small datasets in materials science,” *NPJ Comput Mater*, vol. 4, no. 1, Dec. 2018, doi: 10.1038/s41524-018-0081-z.

[4] L. Wei, X. Xu, Gurudayal, J. Bullock, and J. W. Ager, “Machine Learning Optimization of p-Type Transparent Conducting Films,” *Chemistry of Materials*, vol. 31, no. 18, pp. 7340–7350, Sep. 2019, doi: 10.1021/acs.chemmater.9b01953.

[5] Z. Zhao, Y. Geng, A. Troisi, and H. Ma, “Performance Prediction and Experimental Optimization Assisted by Machine Learning for Organic Photovoltaics,” *Advanced Intelligent Systems*, vol. 4, no. 6, p. 2100261, Jun. 2022, doi: 10.1002/aisy.202100261.

[6] S. Pruksawan, G. Lambard, S. Samitsu, K. Sodeyama, and M. Naito, “Prediction and optimization of epoxy adhesive strength from a small dataset through active learning,” *Sci Technol Adv Mater*, vol. 20, no. 1, pp. 1010–1021, Dec. 2019, doi: 10.1080/14686996.2019.1673670.

[7] R. Yuan *et al.*, “Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning,” *Advanced Materials*, vol. 30, no. 7, Feb. 2018, doi: 10.1002/adma.201702884.

[8] N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, and S. P. Russo, “Machine learning property prediction for organic photovoltaic devices,” *NPJ Comput Mater*, vol. 6, no. 1, Dec. 2020, doi: 10.1038/s41524-020-00429-w.

[9] Y. Cui *et al.*, “Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages,” *Nat Commun*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-10351-5.

[10] X. Du *et al.*, “Elucidating the Full Potential of OPV Materials Utilizing a High-Throughput Robot-Based Platform and Machine Learning,” *Joule*, vol. 5, no. 2, pp. 495–506, Feb. 2021, doi: 10.1016/j.joule.2020.12.013.

[11] M.-H. Lee, “Performance and Matching Band Structure Analysis of Tandem Organic Solar Cells Using Machine Learning Approaches,” *Energy Technology*, vol. 8, no. 3, p. 1900974, Mar. 2020, doi: 10.1002/ente.201900974.

[12] R. Fisher, “The Design of Experiments, ,” *5th ed. (Oliver & Boyd, Oxford, 1949).*.

[13] G.E.P. Box and K.B. Wilson, “On the Experimental Attainment of Optimum Conditions,” *Imperial Chemical Industries*, 1992.

[14] R. N. T. Kackar, “ Off-line quality control, parameter design, and the Taguchi method.,” *J. Qual. Technol.17, 176–188 (1985)*.

[15] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, “Accelerated search for materials with targeted properties by adaptive design,” *Nat Commun*, vol. 7, Apr. 2016, doi: 10.1038/ncomms11241.

[16] I. Izonin, R. Tkachenko, M. Gregus, K. Zub, and N. Lotoshynska, “Input doubling method based on SVR with RBF kernel in clinical practice: Focus on small data,” in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 606–613. doi: 10.1016/j.procs.2021.03.075.

[17] S. Bhatti *et al.*, “Machine learning for accelerating the discovery of high performance low-cost solar cells: a systematic review,” Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.13893

[18] A. Mahmood and J.-L. Wang, “A time and resource efficient machine learning assisted design of non-fullerene small

molecule acceptors for P3HT-based organic solar cells and green solvent selection," *J Mater Chem A Mater*, vol. 9, no. 28, pp. 15684–15695, 2021, doi: 10.1039/D1TA04742F.

[19] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian Process Regression for Materials and Molecules," *Chemical Reviews*, vol. 121, no. 16. American Chemical Society, pp. 10073–10141, Aug. 25, 2021. doi: 10.1021/acs.chemrev.1c00022.

[20] J. Quiñonero, Q. Quiñonero-Candela, C. E. Rasmussen, and C. M. De, "A Unifying View of Sparse Approximate Gaussian Process Regression," 2005.

[21] E. Xing and @ Cmu, "School of Computer Science Probabilistic Graphical Models Learning one Learning one-node GM node GM Reading: Learning Graphical Models Given set of independent samples (assignments of random variables), find the best (the most likely?) Bayesian Network (both DAG and CPDs)," 2009.

[22] E. O. Pyzer-Knapp, G. N. Simm, and A. Aspuru Guzik, "A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials," *Mater Horiz*, vol. 3, no. 3, pp. 226–233, 2016, doi: 10.1039/C5MH00282F.

[23] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans, "Gaussian Process Regression for Optimization."

[24] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," *npj Computational Materials*, vol. 5, no. 1. Nature Publishing Group, Dec. 01, 2019. doi: 10.1038/s41524-019-0153-8.

[25] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active Learning with Statistical Models," 1996.

[26] B. Settles, "Computer Sciences Department Active Learning Literature Survey," 2009.

[27] Jean-Dominique Quinet, « L'Active Learning, une stratégie efficace pour diminuer le coût et le temps du travail préparatoire de vos données. » Onsi, 20, oct. 2020.