Apprentissage des représentations en neuroimagerie : transfert de connaissance à partir de larges jeux de données contrôles vers de petites cohortes cliniques

Benoit Dufumier

université
PARIS-SACLAY

Representation learning in neuroimaging

Transferring from big healthy data to small clinical cohorts

Apprentissage des représentations en neuroimagerie: transfert de connaissance à partir de larges jeux de données contrôles vers de petites cohortes cliniques.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 sciences et technologies de l'information et de la communication (STIC)

Spécialité de doctorat: Mathématiques et informatique

Graduate School : Informatique et sciences du numérique. Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche **BAOBAB** (Université Paris-Saclay, CEA, CNRS), sous la direction d'**Arthur TENENHAUS**, Professeur, la co-direction d'**Edouard DUCHESNAY**, Directeur de Recherche et le co-encadrement de **Pietro GORI**, Maître de conférences

Thèse soutenue à Paris-Saclay, le 16 Décembre 2022, par

Benoit DUFUMIER

Composition du jury

Membres du jury avec voix délibérative

Olivier COLLIOT	Président & Rapporteur
Directeur de recherche, Paris Brain Institute, CNRS	
Ender KONUKOGLU	Rapporteur & Examinateur
Professeur, ETH-Zurich	
Janaina MOURAO-MIRANDA	Examinatrice
Professeur, University College London	
Christos DAVATZIKOS	Examinateur
Professeur, University of Pennsylvania	

THESE DE DOCTORAT

NNT : 2022UPASG093

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Apprentissage des représentations en neuroimagerie: transfert de connaissance à partir de larges jeux de données contrôles vers de petites cohortes cliniques.

Mots clés: Apprentissage profond, Neuroimagerie, Apprentissage des représentations, Troubles psychiatriques

Résumé: La physiopathologie des maladies mentales telles que la schizophrénie et le trouble bipolaire est encore mal comprise, cependant l'émergence de grandes bases de données transdiagnostiques d'images cérébrales offre une occasion unique d'étudier les signatures neuroanatomiques de ces maladies. Le développement de modèles d'apprentissage profonds pour l'imagerie médicale a ouvert la voie à des applications complexes comme la segmentation d'images. Néanmoins, l'applicabilité de telles méthodes aux problèmes de prédiction à l'échelle individuelle à partir d'IRM anatomique reste encore inconnue. Dans cette thèse, nous étudions d'abord la performance des réseaux de neurones actuels en fonction de la quantité de données disponibles. Nous comparons ces performances avec les modèles linéaires régularisés ainsi que les machines à vecteurs de support avec noyau. Nous constatons un problème de sur-ajustement important sur les jeux de données cliniques ainsi qu'une courbe d'apprentissage similaire aux modèles linéaires pour les tailles d'échantillon actuellement accessible en recherche clinique. Nous montrons que cet effet de sur-ajustement est en partie dû au biais induit par les scanners IRM et les protocoles d'acquisition (effet site). Nous proposons une nouvelle solution d'apprentissage des représentations sur de grands jeux de données multi-site d'imagerie de la population saine, basée sur l'apprentissage auto-supervisé par contraste. En transférant ces connaissances à de nouveaux jeux de données cliniques, nous démontrons une amélioration des performances de classification et une plus grande robustesse à l'effet site. Par ailleurs, nous fournissons des garanties théoriques de généralisation de ces modèles pour les tâches de classification. Enfin, pour une meilleure reproductibilité et comparaison des modèles profonds en neuroimagerie, nous introduisons un nouveau jeu de données multi-site à large échelle: OpenBHB. Cette base de données est spécialement conçue pour la prédiction de l'âge cérébrale (tâche supervisée) ainsi que la suppression de l'effet site dans les représentations des modèles profonds. Nous proposong également un défi, accessible en ligne, pour l'apprentissage des représentations avec OpenBHB ainsi qu'une nouvelle méthode pour évaluer le biais dans les représentations des modèles soumis.

Title: Representation learning in neuroimaging: transferring from big healthy data to small clinical cohorts

Keywords: Deep Learning, Neuroimaging, Representation Learning, Transfer Learning, Psychiatric disorders

Abstract: Psychiatry currently lacks of objective quantitative measures to guide the clinician in choosing the right therapeutic treatment. The physio-pathology of mental illnesses such as schizophrenia and bipolar disorder is still poorly understood but the emergence of large-scale neuroimaging transdiagnostic datasets gives a unique opportunity for studying the neuroanatomical signatures of such diseases. While Deep Learning (DL) models for medical imaging unlocked unprecedented applications such as image segmentation, its applicability to single-subject prediction problems with neuroanatomical MRI yet remains limited. In this thesis, we first study the current performance and scaling trend of DL models, for several architectures representative of the recent progression in computer vision, as compared to regularized linear models and Kernel Support Vector Machine. We found a high over-fitting issue on clinical data-sets and a similar scaling trend with linear models, for current accessible sample size in clinical research. This over-fitting effect was also due to the bias induced by MRI scanners and acquisition protocols. To tackle the sample size issue, we propose a new method to learn a representation of the healthy population brain anatomy on large multi-site cohorts with neural networks using contrastive learning, an innovative self-supervised framework. When transferring this knowledge to new datasets, we demonstrate an improvement in classification performance of patients with mental illnesses. We provide a theoretical framework grounding these empirical results and we show good generalization properties of the model for downstream classification tasks with weaker hypothesis than in the literature. Moreover, as an advancement towards debiased deep models and reproducibility in neuroimaging, we introduce a new large-scale multi-site dataset, OpenBHB, for brain age prediction and site de-biasing as well as a permanent challenge focused on representation learning. We offer three pre-processing to study brain anatomical surface, geometry and volume inside T1 images as well as a novel way to evaluate the bias in model's representation.

À Javi, pour ton éternel soutien

Acknowledgements

My PhD is the result of exciting collaborations, open discussions as well as supportive friends and a loving family. It is then obvious that they should occupy a part of this text. These acknowledgements will be sometimes in French or in English, depending on the targeted reader.

J'aimerais commencer ces remerciements avec mon premier encadrant (et directeur) de thèse, Arthur Tenenhaus. Tu as été le premier à m'avoir introduit au monde de la recherche, un monde si particulier par son exigence mais également si enrichissant, professionnellement parlant bien sûr, mais également personnellement... Cela m'a donné l'envie de continuer plus tard à NeuroSpin avec mon deuxième encadrant, Edouard Duchesnay, à qui je dois un soutien et une écoute constante à travers cette thèse marquée par deux années de covid. Je n'oublierai pas ces longues discussions, sous la pluie, sous la neige ou sous le soleil saclaysien durant les courses hebdomadaires du vendredi midi! Enfin, je remercie très sincèrement Pietro Gori, mon troisième et dernier encadrant. Tu m'as toujours aiguillé, épaulé et soutenu dans l'ensemble de mes travaux et tu as sû m'apprendre à mener ces projets à leur terme, bon gré mal gré. Tu m'as également permis de découvrir toutes les facettes d'un chercheur, tantôt encadrant, tantôt prof, tantôt écrivain, tantôt développeur mais toujours curieux et ouvert à de nouveaux projets et de nouvelles idées. Merci à tous les trois pour votre confiance et votre aide tout au long de ce voyage qui m'a aidé à construire mon projet.

Then I would like to sincerely thank all the jury of my thesis (Ender Konukoglu, Janaina Mourao-Miranda, Christos Davatzikos and Olivier Colliot) for their careful evaluation of my manuscript and their interest for this PhD. I am very glad for all their positive comments that really moved me during my defense. I will keep very good memories of this moment and I hope it will foster new opportunities for our mutual projects.

Bien sûr, l'environnement de recherche est l'un des facteurs les plus importants pour mener à bien une thèse et je voudrais remercier chaleureusement l'équipe BAOBAB dans son ensemble. En particulier, merci à tous les doctorant(e)s et post-docs du groupe de lecture de Deep Learning à NeuroSpin: Anaïs, Anas, Aymeric, Bastien (les deux), Benjamin, Chaithya, Chloë, Clément, Denise, François, Joël, Ilias, Guillaume, Kumari, Pierre, Sara, Simon et tant d'autres dont je m'excuse déjà de ne pas les citer (ce groupe s'est tellement élargi !). Ne m'en voulez pas, je n'oublie évidemment pas les membres fondateurs de ce groupe: Louise, Alex et Zac avec qui j'ai adoré partagé tous nos moments (y compris à l'extérieur de NeuroSpin pour élaborer nos plans machiavéliques). Je fais une dédicace toute particulière à Coco et Robin pour reprendre vaillamment ce flambeau et pour continuer à faire vivre ce groupe. Je remercie aussi particulièrement Carlo comme premier membre extérieur VIP dans ce groupe.

Je suis aussi reconnaissant à notre (plus petite) équipe signature avec qui j'ai partagé toutes les aventures de la thèse: Antoine, Julie (toujours souriante) puis Robin (le plus endurant tant en athlétisme qu'en alcool), Ilias (parti trop tôt) puis Loic (le plus beau) et enfin Pierre et Sara (qui m'a tant appris sur l'Iran même si le temps m'a manqué pour en discuter plus avant). Et puis je n'oublie pas ses deux grandes soeurs Brainomics et feu NAO (un peu plus âgées) avec une dédicace toute particulière à Louise, Coco et Cyril pour les moments en salle de pause (c'est-à-dire nos bureaux...) et toutes les discussions (philosphiques, informatiques, mathématiques, gastronomiques, métaphysiques, random, improbables, personnelles et j'en passe) qui font également parties de cette thèse. Je remercie évidemment Jean-François pour tes conseils ainsi que pour les références toujours brillantes et inspirantes qui m'ont aidé à me poser les bonnes questions. Merci à Joël pour toutes nos discussions contrastées et ta bonne humeur. Enfin, je remercie Vincent pour avoir toujours poussé à la collaboration entre nos équipes.

Les vendredis midi ont embelli ma semaine durant ces années de thèse grâce aux éternels coureurs Neurospiniens, Nico et François, que je remercie avec joie. Egalement, François, merci pour m'avoir introduit à Unicog même si je n'ai pas su pleinement profiter de cette opportunité (la partie n'est que remise). Le sport a toujours été source d'équilibre pour moi et c'est pourquoi je me dois de remercier tous les grimpeurs avec qui j'ai eu le plaisir de partager des journées

entières à Fontainebleau ou Arkose, je pense en particulier à Gaston, Félix, Séb, DiLo, Alex, Coco, Matt, Laure, Pierre, Solène et bien entendu Vincent. Nos aventures ardéchoises resteront un moment unique, où j'ai aussi découvert qu'en recherche, les vacances sont fonction des rebuttals en non l'inverse...

Mon doctorat a été partagé durant deux années entre NeuroSpin et Télécom. Ainsi, je remercie l'ensemble des doctorants de l'équipe IMAGES et les membres du groupe de travail en Deep Learning à Télécom pour leur motivation, leur engagement dans ce groupe et leurs idées qui m'ont largement inspiré pour la co-création de son homologue Neurospinien. Of course, I would like to greatly thank Carlo (again) for all the inspiring discussions about AI we had, your incredible intuition and ideas that made my third year a quite unique journey. I won't forget about this crazy night at Télécom for our first joint submissions at NeurIPS, and I am looking forward crossing your path again in the future.

Merci à Serge pour ton aide lors de la soutenance et à Véronique pour être présente dans les moments importants.

A mes amis de toujours, Vincent et Dylan qui, sans aucun doute, ont été et resteront à mes côtés dans les meilleurs mais aussi les pires moments.

A mes parents, qui m'ont toujours épaulé et soutenu dans tous mes choix. Votre amour et votre confiance sont les plus précieux à mes yeux, ils me permettent d'avancer et de me construire. Cette thèse n'en est qu'une illustration et vous en resterez les co-auteurs inapparents.

Enfin, je termine en laissant quelques mots à Javi, un être incroyable avec qui je partage ma vie depuis déjà cinq ans. Cette thèse n'aurait certainement pas abouti sans toi: le doctorat est déjà un moment éprouvant mais les deux années de covid nous ont fait rentrer dans une période d'isolement assez unique. Ton aide, tes conseils et tes idées m'ont fait avancer dans ce voyage tortueux et son aboutissement n'en est que le fruit. Je suis heureux et fier d'être avec toi, de regarder dans la même direction vers l'avenir et de construire ensemble notre projet. Et puisque l'essentiel est indicible, je m'arrête là en espérant le laisser visible avec mon coeur.

Contents

Résumé

Introduction

La neuroimagerie permet d'étudier le cerveau humain afin de comprendre comment ce système biologique peut accomplir des tâches cognitives de haut niveau (langage, mémoire, attention, raisonnement, perception et émotion) mais aussi comme outil de diagnostic pour le clinicien. Au début de l'imagerie cérébrale, cette technique a permis plusieurs avancées pour identifier des lésions ou des tumeurs cérébrales (par exemple, l'angiographie cérébrale mise au point en 1927), mais elle nécessitait des interventions dangereuses et douloureuses pour le patient. Par la suite, de nouvelles techniques d'imagerie ont été développées, dont l'imagerie par résonance magnétique (IRM), basée sur la propriété physique du proton à l'intérieur des molécules d'eau sous un champ magnétique élevé. Cette technique non invasive fournit des informations sur la structure du cerveau (comme les cartes de connectivité cérébrale entre les régions avec l'IRM de diffusion) et sur l'activité cérébrale (via des mesures indirectes du flux sanguin avec l'IRM fonctionnelle par exemple).

L'apport de l'IRM pour l'étude des maladies psychiatriques. Pour les troubles psychiatriques tels que la schizophrénie, le trouble bipolaire ou les troubles du spectre autistique (TSA), il n'existe actuellement aucun biomarqueur objectif et quantitatif dans le cerveau (et par extension, aucun test clinique) pour guider le clinicien dans le choix d'une stratégie thérapeutique ciblée à l'échelle individuelle. Le diagnostic de ces maladies repose uniquement sur des entretiens cliniques et des questionnaires permettant de rapporter des symptômes qui sont ensuite classés selon le Manuel diagnostique et statistique (DSM). Les recherches antérieures ont principalement étudié ces troubles à l'échelle du groupe en identifiant les caractéristiques anormales du cerveau dans un groupe de patients par rapport à des sujets sains au moyen de tests statistiques. Cette approche a permis de découvrir plusieurs biomarqueurs pertinents pour les troubles cérébraux (comme des anomalies de connectivité dans le système limbique-striatal pour les TSA [199] ou des connexions fonctionnelles plus élevées entre les régions pour la schizophrénie [157]), mais son application au diagnostic/pronostic clinique est difficile, principalement dû au manque de pouvoir discriminant des biomarqueurs trouvés à l'échelle du groupe [9], ce qui empêche leur utilisation au niveau individuel.

L'apprentissage automatique pour la médecine de précision. Les modèles d'apprentissage automatique (ML) offrent une solution attrayante pour aborder la prédiction individuelle à partir de données IRM. Au lieu de considérer un effet statistiquement significatif à l'échelle du groupe, le modèle est entraîné à prédire *un état clinique par sujet*. Une fois entraîné, le modèle peut ensuite prédire cet état clinique à partir de nouvelles entrées par extrapolation. En outre, l'émergence de jeux de données à grande échelle provenant de plusieurs consortiums internationaux (comme le Human Connectom Project [286], ABIDE [79, 80], UKBioBank [37]) rendent possible l'entraînement de ces algorithmes de plus en plus complexes qui peuvent *"aider à découvrir de nouveaux mécanismes causaux et conduire à la génération de nouvelles hypothèses"* [257] (i.e. découverte de biomarqueurs) ainsi qu'aider le clinicien à choisir le bon traitement face à un patient souffrant de plusieurs maladies plausibles.

La piste priviligiée: l'apprentissage des représentations par transfert. En partant du constat que les jeux de données à large échelle de contrôles sains sont maintenant disponibles alors que les cohortes de patients avec troubles psychiatriques *homogènes* (i.e. scannés avec le même scanner/protocole, ayant pris les mêmes traitements et avec le même diagnostic) sont, et seront dans un futur proche, à petite échelle, nous nous demandons: pouvons-nous changer le paradigme supervisé traditionnel en ML pour exploiter ces larges jeux de données contrôles pour la prédiction des troubles cérébraux? Dans cette thèse, nous étudions les modèles d'apprentissage profond (DL) hiérarchique afin d'apprendre la représentation des données d'imagerie cérébrale de sujets sains, et de découvrir les signatures neuroanatomiques discriminantes des sujets malades au sein de petites cohortes cliniques.

Potentiel et limites de l'apprentissage supervisé des représentations pour la neuroimagerie

Dans ce chapitre, nous avons étudié les principales propriétés des modèles supervisés de DL sur des données d'imagerie cérébrale anatomique. Pour mener à bien notre analyse, nous avons d'abord rassemblé une grande collection d'images cérébrales par le biais de diverses initiatives de partage, ce qui nous a permis d'obtenir un vaste ensemble de données multi-sites. Il comprend notamment des patients atteints de schizophrénie, de troubles bipolaires et d'autisme, mais aussi une grande base de sujets sains.

À partir de ce jeu de données, nous avons montré que les modèles à l'état de l'art en DL sont aussi performants que les modèles linéaires régularisés pour les tailles d'échantillon clinique actuelles sur les tâches de classification des troubles mentaux. Ils ont tendance à surajuster rapidement, notamment sur le bruit associé au site d'acquisition, ce qui les empêche– entre autres– d'extraire des motifs géométriques discriminants (par exemple les plis corticaux) enfouis dans les images IRM brutes. Nous avons observé ce comportement à plusieurs reprises en analysant leurs performances sur des ensembles de tests inter-sites externes et cela met

en lumière un biais important dans les jeux de données de neuroimagerie actuels qui sera certainement amplifié au fur et à mesure que d'autres initiatives verront le jour. Il est intéressant de noter que que les DNN étudiés restent biaisés même lorsqu'ils sont entraînés sur des données à large échelle ($N = 10$k) pour la prédiction du phénotype, ce qui suggère que "tout n'est pas une question de taille de données", comme cela a été illustré sur la maladie d'Alzheimer par Varoquaux et Cheplygina [290].

À partir de cette analyse, nous avons étudié l'augmentation des données comme technique de régularisation ainsi que les techniques de débiaisage basées sur les données (telle que l'harmonisation multi-site) pour les réseaux de neurones. Nous n'avons trouvé aucune amélioration pour les applications cliniques ciblées, ce qui suggère que les augmentations actuelles conçues à partir de la perception humaine doivent être repensées pour l'imagerie cérébrale.

Enfin, comme l'envisagent Bzdok, Floris et Marquand [40], la modélisation de la variabilité biologique et de l'incertitude méthodologique par le biais de la théorie bayésienne est requise pour l'analyse de l'IRM cérébrale afin d' "*aller au-delà des affirmations binaires sur l'existence ou la non-existence d'un effet et fournir des estimations de crédibilité autour de tous les paramètres du modèle en jeu, ce qui permet ainsi des prédictions par sujet avec des intervalles d'incertitude rigoureux.*". Par conséquent, dans la dernière section, nous avons utilisé les travaux récents sur les réseaux neuronaux bayésiens pour modéliser les incertitudes aléatoires et épistémiques dans les DNN, en replaçant les techniques standard de Dropout et de Deep Ensemble dans ce cadre. Nous montrons notamment une amélioration significative de la calibration et des performances sur toutes les tâches de classification des troubles psychiatriques avec des DNN largement sur-paramétrés. Ce travail souligne l'importance de la modélisation de l'incertitude épistémique et ouvre de nouvelles voies pour le développement de nouvelles approximations variationnelles de la distribution postérieure du réseau.

Apprentissage non-supervisé des représentations pour la neuroimagerie

Dans le chapitre précédent, nous avons cherché à découvrir la capacité de représentation des DNN sur données d'imagerie cérébrale dans un contexte entièrement supervisé pour discriminer les patients des sujets sains. L'une des principales limites de cette approche tient au besoin toujours croissant de (très) grands jeux de données pour obtenir une convergence satisfaisante. Cela a été illustré dans le chapitre précédent sur les tâches de classification pour détecter des troubles psychiatriques mais aussi de regression du phénotype (comme l'âge), où les réseaux neuronaux ne convergeaient pas vers de meilleures solutions que les modèles linéaires, pour une taille d'échantillon inférieure à 1000.

De grandes initiatives pour imager la population , telles le Human Connectom Project [286] (lancé en 2010) ou UKBioBank [37] (lancé en 2006 qui a déjà imagé près de 100 000 sujets

au Royaume-Uni) – axées principalement sur la population saine – permettent maintenant le développement de nouveaux outils d'IA pour modéliser le développement normal du cerveau humain tout au long de la vie (de l'enfance à la vieillesse). Cette nouvelle ressource permet de modéliser avec précision la variabilité biologique inhérente du cerveau sain (par exemple, associée aux informations phénotypiques ou génotypiques telles que l'âge, le sexe ou le score polygénique) comme une variété dans un espace de faible dimension. De ce point de vue, les cerveaux pathologiques (par exemple, ceux des sujets atteints de schizophrénie ou de troubles bipolaires, qui présentent des schémas cérébraux corticaux anormaux par rapport au groupe sain) peuvent être considérés comme une déviation orthogonale à l'espace vectoriel tangent de son "jumeau sain" (non observé), situé sur cette variété (comme l'a illustré Aglinskas et al. dans un article récent de Science [4] consacré à l'autisme).

Dans ce chapitre, nous étudions comment modéliser ce type de variété de faible dimension de la population saine en utilisant des modèles auto-supervisés basés sur l'apprentissage par contraste (CL). Ces modèles discriminatifs présentent plusieurs avantages par rapport à leurs homologues génératifs (tels que le VAE [174] ou le GAN [116]) : ils ne nécessitent pas une génération à l'échelle du pixel, exigeante en calcul et qui reste une tâche difficile, ils sont faciles à entraîner et ils ne modélisent pas explicitement le processus de génération des données mais plutôt une approximation de son inverse (depuis l'espace observable vers l'espace latent non observé [323]). Nous validons les modèles développés dans ce chapitre sur plusieurs cohortes cliniques incluant des patients atteints de schizophrénie, de troubles bipolaires, d'autisme mais aussi d'Alzheimer, couvrant ainsi un large spectre des troubles psychiatriques et neurodénératifs.

Dans la première partie, nous présentons la formulation originale du CL pour l'apprentissage des représentations visuelles [52, 124, 211] du point de vue de la théorie de l'information et nous présentons ses deux principales implémentations avec MoCo [136] et SimCLR [52]. Comme première contribution originale, nous décrivons comment des informations phénotypiques auxiliaires telles que l'âge du sujet peuvent être exploitées pour pour mieux apprendre les motifs caractéristiques et discriminants de la population saine vis-à-vis des cerveaux malades. Ce cadre étend notamment l'apprentissage par contraste supervisé au cas faiblement supervisé en utilisant une nouvelle fonction de similarité entre les variables auxiliaires. Nous étudions également les composantes critiques de ce modèle, telle que l'augmentation des données et la taille des lots, et leur impact sur la représentation finale du modèle.

Dans la deuxième partie, nous fournissons un cadre théorique à l'apprentissage par contraste. Sur la base de cette analyse, nous nous demandons si le module d'augmentation des données (composante critique dans les modèles actuels de CL) peut être partiellement retiré pour l'apprentissage des représentations en imagerie médicale. Pour ce faire, nous développons une nouvelle théorie basée sur l'intégration d'une fonction noyau entre images sur un espace à noyaux reproduisants, vue comme a priori durant l'apprentissage. Nous montrons notamment que les modèles génératifs ou les variables auxiliaires associées aux images peuvent définir un tel

a priori. Nous démontrons des nouvelles bornes sur le risque supervisé avec moins d'hypothèses que la littérature actuelle et nous explorons et validons cette nouvelle approche sur des données d'imagerie IRM et de radiographie thoracique.

OpenBHB: un nouveau défi pour l'apprentissage supervisé et le débiaisement

Avec l'émergence croissante de nouvelles ressources multi-sites à large échelle pour la neuroimagerie, nous anticipons l'émergence de modèles profonds pour l'apprentissage des représentations supervisé. Cependant, comme nous l'avons vu dans le chapitre 2, ces données d'imagerie sont souvent collectées avec des scanners et des protocoles d'acquisition différents. Ces disparité influencent fortement la qualité des images et induisent un biais important dans les modèles d'apprentissage automatique, phénomène bien décrit dans le chapitre 2. Comme l'a supposé D. Bzdok [38], l'hétérogénéité inter-site peut expliquer pourquoi, de manière contre-intuitive, il a été signalé à plusieurs reprises que les performances des modèles prédictifs diminuent à mesure que les données neuroscientifiques disponibles augmentent [306].

Ce chapitre est consacré à la résolution de ce problème. Nous présentons d'abord une nouvelle ressource d'IRM cérébraux à large échelle – OpenBHB – accessible librement par tous, ainsi qu'un nouveau défi pour la prédiction de l'âge biologique avec suppression de l'effet site, considéré comme une tâche de débiaisage. L'estimation précise de l'âge biologique à partir de l'imagerie cérébrale reste un défi important pour la communauté qui peut permettre la découverte de nouveaux biomarqueurs (par exemple en utilisant la différence entre âge biologique et chronologique comme proxy pour la caractérisation de l'accélération du veillissement cérébral chez des sujets malades). OpenBHB est assez unique par sa taille (comprenant $N > 5$k sujets) et son hétérogénéité (71 centres d'acquisition répartis dans le monde entier sur 3 continents - Asie, Amérique du Nord et Europe). Ce jeu de données est centré sur la population saine et il est doté de pipelines de pré-traitement standardisés pour l'analyse IRM de surface et de volume. Dans une première partie, nous présentons d'abord les propriétés statistiques d'OpenBHB avant de décrire le défi actuellement disponible sur la plateforme RAMP. Ce défi introduit des métriques dérivées de la représentation des modèles soumis (en particulier via l'évaluation linéaire [5]). Ces métriques quantifient à la fois le biais associé aux sites mais aussi les performances de généralisation inter-site des modèles pour la prédiction de l'âge biologique cérébral. Dans une seconde partie, nous présentons les premières expériences et résultats des modèles profonds entraînés sur plusieurs modalités d'IRM (volumique incluant la densité de matière grise et surfacique incluant l'épaisseur corticale, la surface, la courbure locale, etc.). Nous comparons les performances de ces modèles avec l'état de l'art pour l'harmonisation multi-site, à savoir ComBat [101]. Nous ouvrons enfin des perspectives avec l'apprentissage par contraste en proposant un nouveau terme de régularisation dans l'objectif à optimiser, qui inclut le biais associé au site d'acquisition.

Introduction

Neuroimaging allows to investigate the human brain in order to understand how this biological system can perform high-level cognitive tasks (language, memory, attention, reasoning, perception and emotion) but also as a diagnostic tool for the clinician. In the early days of brain imaging, this technique led to several breakthroughs for identifying brain lesions or brain tumors (e.g. with cerebral angiography developed in 1927) but it required dangerous and painful interventions for the patient (e.g. involving injection of filtered air in ventricular system for pneumoencephalography or injection of contrast agent for angiography). Later on, new imaging techniques have been developed, among which Magnetic Resonance Imaging (MRI) in the 70's based on the physical property (spin) of the proton inside water molecules under a high magnetic field. This non-invasive technique provides information about brain structure (e.g. brain connectivity maps between regions with diffusion MRI) and brain activity (e.g. by indirect measures through blood-flow with functional MRI). It can be used as a diagnosis tool for clinicians, in particular for conditions involving the central nervous system such as cerebrovascular disease, epilepsy or demyelinating disorders. Additionally, it can finely assess the degree of brain injury (e.g. after a stroke) and identify vascular lesions responsible of a specific disorder (e.g. ischemic stroke).

Brain MRI is thus an imaging technique that provides brain observations that we cannot see with our naked eyes and it has fostered our understanding of the human brain for both neuroscience and clinical applications in the last 50 years. Nonetheless, for brain disorders such as schizophrenia, bipolar disorder or autism spectrum disorders (ASD), there is currently no objective and quantitative biomarkers in the brain (and by extension, clinical tests) available to guide the clinician in choosing a therapeutic strategy. The diagnosis of such diseases is only based on clinical interviews and questionnaires that allow a reporting of symptoms that are then classified based on the Diagnostic and Statistical Manual (DSM). Past research has mainly investigated these disorders at the group-level by identifying brain abnormal features in a group of patients vs healthy subjects through statistical tests. This approach unveiled several relevant biomarkers for brain disorders (e.g. connectivity abnormalities in the limbic-striatal system for ASD [199] or higher functional connections between regions for schizophrenia [157]), yet its translation to clinical diagnosis or prognosis is difficult. One main reason that explains this difficulty is the lack of discriminative power from the biomarkers found at the group-level [9],

preventing its adoption at the individual level.

Machine learning models offer an appealing solution for tackling subject-level prediction from brain imaging data. Instead of considering a statistically significant effect at the group level (for a fixed p-value), the model is trained to predict a clinical status (diagnosis, prognosis or other phenotype) from a single entry. Once trained, the model can then predict this clinical status from new arriving entries by extrapolation. As noted by Bzdok and Meyer-Lindenberg [39], "*machine learning and classical statistics do not judge data on the same aspects of evidence: an observed effect assessed to be statistically significant by a p-value does not in all cases yield a high prediction accuracy in new, independent data, and vice versa*". Additionally, the emergence of large-scale datasets from several international consortium (e.g. Human Connectom Project [286], ABIDE [79, 80], UKBioBank [37]) allows to train increasingly complex ML algorithms that can *"help uncover potential new causal mechanisms and lead to the generation of new hypotheses"* [257] (*i.e.,* biomarker discovery) as well as help the clinican in choosing the right treatment when facing a patient with multiple plausible diseases.

Nonetheless, such large emerging datasets come with several challenges. First, they are often transdiagnostic, gathering patients with various medication histories and symptoms severity which highly reduce the number of homogeneous patients, with no comorbidity, for a given brain disorder. Small sample size is a major issue for ML models as it easily leads to over-fitting on training data: the model memorizes each training image by learning spurious patterns (*e.g.,* associated to noise-specific features) with no capacity of generalization on new, unseen images. It also leads to high error-bar for its predictions [289], which can bias the scientific community towards over-optimitic results [290] (*e.g.,* through "over-fitting by observer" when a cross-validation model is cherry-picked by the researcher). On the other hand, we should emphasize that large cohorts of healthy controls with no history of mental disorders are now easily available. Second, international consortium (*e.g.,* ABIDE [79, 80] or ABCD Study) often gather brain MRI acquired on various acquisition sites with different scanners and acquisition settings (magnetic field, imaging sequence, etc.), introducing a strong bias in the resulting images that may heavily hurt the generalization performance on external data coming from never-seen sites.

Considering that large-scale datasets of healthy controls are now available while homogeneous cohorts of patients (*e.g.,* scanned with the same scanner/protocol with same medication history and diagnosis) are, and will be in the near future, small-scale, we ask: can we change the traditional supervised paradigm in ML to leverage these large datasets of healthy controls for single-subject prediction of brain disorders ? In this thesis, we investigate deep learning models in order to learn representation of brain imaging data, in a layer-wise manner, and to discover the hidden structure in the data along with the relevant axis of variations that allows to discriminate a patient from the healthy population. As a result, our approach in this thesis complies with the recommendation provided by the Research Domain Criteria (RDoC)

initiative [153] that wants to "*better understand basic dimensions of functioning that span the full range of human behavior from normal to abnormal*". Deep models are particularly well suited for learning representations [30] both in an unsupervised and supervised setting on vision tasks. Very recent developments in computer vision have shown that they can learn from a broad variety of images at a very large scale (*e.g.,* from million to billion images [117]) with very good transfer performance on downstream classification tasks. It currently leads to a change in paradigm in AI where standard supervised models are replaced by *foundation models* [32], pre-trained on such large-scale datasets and fine-tuned on specific downstream tasks with transfer learning. We argue such shift in paradigm is fully in line with the approach we follow here. We believe that our study of deep models pre-trained on large-scale neuroimaging datasets of the healthy population is a first step towards precision medicine in psychiatry and it will foster the development of innovative, reproducible and open models, at the intersection between neuroimaging, deep learning and computer vision. Models

Objectives

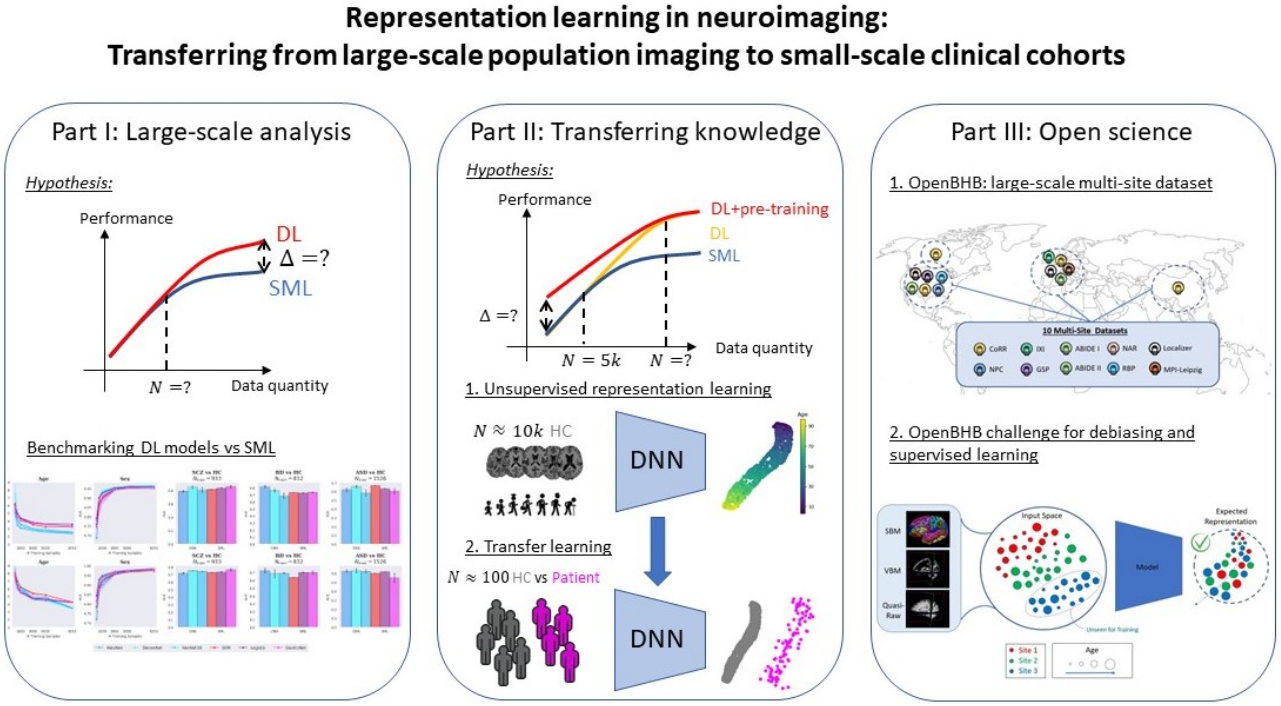


Figure 1: In this thesis, we study representation learning models for single-subject prediction using brain anatomical imaging. First, we perform a large-scale analysis in a supervised context and we study the learning curves of Deep Learning (DL) models against Standard Machine Learning (SML). Then, we present our main paradigm based on transfer learning from a large-scale population imaging dataset of healthy controls to small-scale clinical cohorts in order to discriminate brain disorders from controls using previous knowledge. Finally, we introduce a new benchmarking resource–OpenBHB–along with a challenge to perform supervised representation learning on brain age prediction while keeping a debiased representation, independent from acquisition sites.

This thesis studies the potential of deep learning models for single-subject prediction from

neuroimaging data. Our main goal is to provide a new paradigm to exploit large MRI datasets of the healthy population with deep learning tools in order to improve discrimination performance of brain disorders and ultimately i) help the clinician for choosing better treatment; ii) discover new individual signatures (*i.e.,* patterns) of such highly heterogeneous disorders, hopefully at a very early stage. The main brain disorders considered in this thesis are schizophrenia, bipolar disorder and Autism Spectrum Disorders (ASD) with also possible application to neurological disorder such as Alzheimer's disease. Throughout this thesis, our main experiments will be focused on structural MRI data, which provide information about whole-brain anatomy at the millimetric level.

As a result, the research questions we would like to address in our study are:

1. Can we learn deep non-linear representation from brain imaging data for single-subject prediction of brain disorders and phenotype ? How do these models perform compared to vanilla linear models ? What regularization strategy is best suited ? Do deep models benefit from raw data ? Can we quantify and improve predictive uncertainty to improve downstream representation/performance ? We study these questions by gathering a large-scale transdiagnostic dataset and we show comparable performance between deep and linear models for brain disorder prediction with medium-scale datasets but better scaling trend in the large-scale regime for phenotype prediction (age). Improving predictive uncertainty leads to better deep representation, outperforming the linear baseline also on brain disorder prediction.

2. Can we benefit from large-scale brain images of healthy controls to perform downstream classification of patients with mental disorders with transfer learning ? Are self-supervised algorithms relevant for pre-training such models ? Can we provide theoretical guarantees on downstream task performance ? We develop new self-supervised models capable of integrating auxiliary information from healthy controls (such as phenotype) to shape the representation space. When transferred, we demonstrate better generalization performance on cross-site clinical cohorts, largely outperforming all state-of-the-art models. We analyze theoretically the models and prove generalization guarantees under milder assumption than the current literature.

3. Are deep models representation biased by acquisition scanner ? Can we debias this representation in a open and reproducible way ? We systematically show a bias in deep representations when performing phenotype and brain disorders prediction tasks, leading to poor generalization capacity on cross-site images. We present a new challenge designed for representation learning and debiasing along with an openly accessible large-scale dataset to tackle brain age prediction with site-effect removal.

Thesis organization

This thesis is at the intersection between several fields, notably neuroimaging, machine learning and psychiatry. We start by presenting the data and models used throughout this manuscript in Chapter 1, in particular for brain imaging analysis with machine learning models.

The next three Chapters present our main contributions. In Chapter 2, we explore the discrimination capacity of deep learning models in a supervised context on single-subject prediction tasks using brain imaging. We study several architectures and compare their performance with linear models using both medium and large-scale data volume. We also present several techniques for quantifying predictive uncertainty and ultimately demonstrate their benefit for deep models.

Accounting for the over-fitting issue observed in the medium-scale data regime, we present a new paradigm for deep models in Chapter 3, based on transfer learning. We develop and mathematically analyze several self-supervised techniques for pre-training such models on large healthy datasets, based on contrastive learning, and we demonstrate good generalization performance on several brain disorder classification tasks. We also bridge the gap between generative and discriminative models for pre-training, in particular in the neuroimaging context.

Based on our previous analysis on large-scale multi-site datasets in Chapter 2, we present in Chapter 4 a new challenge for debiasing deep model representation of brain scans from site-related effects while preserving biological variability associated to age. Along with this challenge, we introduce OpenBHB, the first large-scale dataset openly accessible to tackle this problem in neuroimaging. We perform an in-depth analysis of OpenBHB and show first baseline results.

Finally, we conclude this thesis by summarizing our main contributions and findings during this PhD, and we provide several future axis of research for deep representation learning in neuroimaging.

Contributions

This PhD has led to several publications in peer-reviewed journals and international conferences (listed below).

Journal articles

(J1) **Deep Learning Improvement over Standard Machine Learning in Anatomical Neuroimaging comes from Transfer Learning**, B. Dufumier, P. Gori, J. Victor, R. Louiset, J-F Mangin, A. Grigis, E. Duchesnay, *Submitted to NeuroImage*, 2023

(J2) **OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing**, B. Dufumier, A.Grigis, J. Victor, C. Ambroise, V. Frouin, E. Duchesnay,

NeuroImage, 2022

International conference articles

(C1) **Contrastive Learning with Continuous Proxy Meta-Data for 3D MRI Classification**, B. Dufumier, P. Gori, J. Victor, A. Grigis, E. Duchesnay et al., *MICCAI*, 2021

(C2) **Rethinking Positive Sampling for Contrastive Learning with Kernel**, B. Dufumier, C. A. Barbano, R. Louiset, E. Duchesnay, P. Gori, *Submitted to ICML 2023*

(C3) **Supervised Contrastive Learning for Debiasing**, C. A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, P. Gori, *ICLR*, 2023

(C4) **Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels**, B. Dufumier, P. Gori, J. Victor, A. Grigis, E. Duchesnay, *NeurIPS Workshop on Medical Imaging Meets NeurIPS*, 2021

(C5) **Contrastive learning for regression in multi-site brain age prediction**, C. A. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, Pietro Gori, *ISBI*, 2023

Other collaborations

(C6) **UCSL: A Machine Learning Expectation-Maximization framework for Unsupervised Clustering driven by Supervised Learning**, R. Louiset, P. Gori, B. Dufumier, J. Houenou, A. Grigis, E. Duchesnay, *ECML-PKDD*, 2022

(C7) **Unsupervised Representation Learning of Cingulate Cortical Folding Patterns**, J. Chavas, L. Guillon, M. Pascucci, B. Dufumier, D. Rivière, J-F Mangin, *MICCAI*, 2022

(C8) **Detection of abnormal folding patterns with unsupervised deep generative models**, L. Guillon, B. Cagna, B. Dufumier, J. Chavas, D. Rivière, J-F Mangin, *MICCAI MLCN Workshop*, 2021

Chapter 1

Related works

Contents

This chapter introduces the basis for all analysis and methods developed in the rest of this work. In particular, this thesis focuses on deep learning tools for the analysis of structural neuroimaging data. As a result, we first define what are the data we are manipulating in the context of single-subject prediction and its utility for discriminating brain disorders ("what" and "why"). Then, we present the basic notions of machine learning (starting from simple linear models to highly non-linear deep neural networks) in the unsupervised and supervised context. Focusing on Convolutional Neural Networks for image analysis, we describe the different architectures and their evolution that led to current state-of-the-art models for vision applications. Finally, we present recent applications using whole-brain imaging data for psychiatric condition classification.

1.1 Anatomical brain MRI for brain disorders understanding

Brain MRI offers a non-invasive way to investigate the brain. This imaging technique allows to study brain anatomy, structure and function *in vivo* through various physical principles. This thesis will mostly focus on brain anatomical data, so we quickly draw an overview of human brain anatomy before presenting the anatomical features measured with MRI and the main pre-processing techniques used to analyze such data. We conclude by presenting some findings on schizophrenia and bipolar disorder showing patterns of abnormalities using anatomical data.

1.1.1 Anatomical features

Brain anatomy. The brain is part of the central nervous system (along with spinal cord) and it can be decomposed into three areas: the brain stem, the cerebellum ("little brain") and the cerebrum (the largest part). Cerebellum and brain stem are responsible for autonomic processes (*e.g.,* heart rate and breathing) along with balance and coordinate movements. Cerebrum is responsible for high-level cognitive tasks such as information processing, decision-making, memory, emotions and learning. It can be decomposed into gray matter (cerebral cortex) and white matter at its center.

The cerebral cortex is divided into four lobes (see Fig. 1.1), each one of them related to specific functions, among others:

- the **frontal lobe**, in charge of reasoning and decision-making. It notably includes Broca's area which is associated with language processing;
- the **parietal lobe**, responsible for sensory integration, visuo-spatial processing, recognition, and navigation;
- the **occipital lobe**, involved with visual processing;
- the **temporal lobe**, responsible for short and long term memory, language processing, and emotion association.

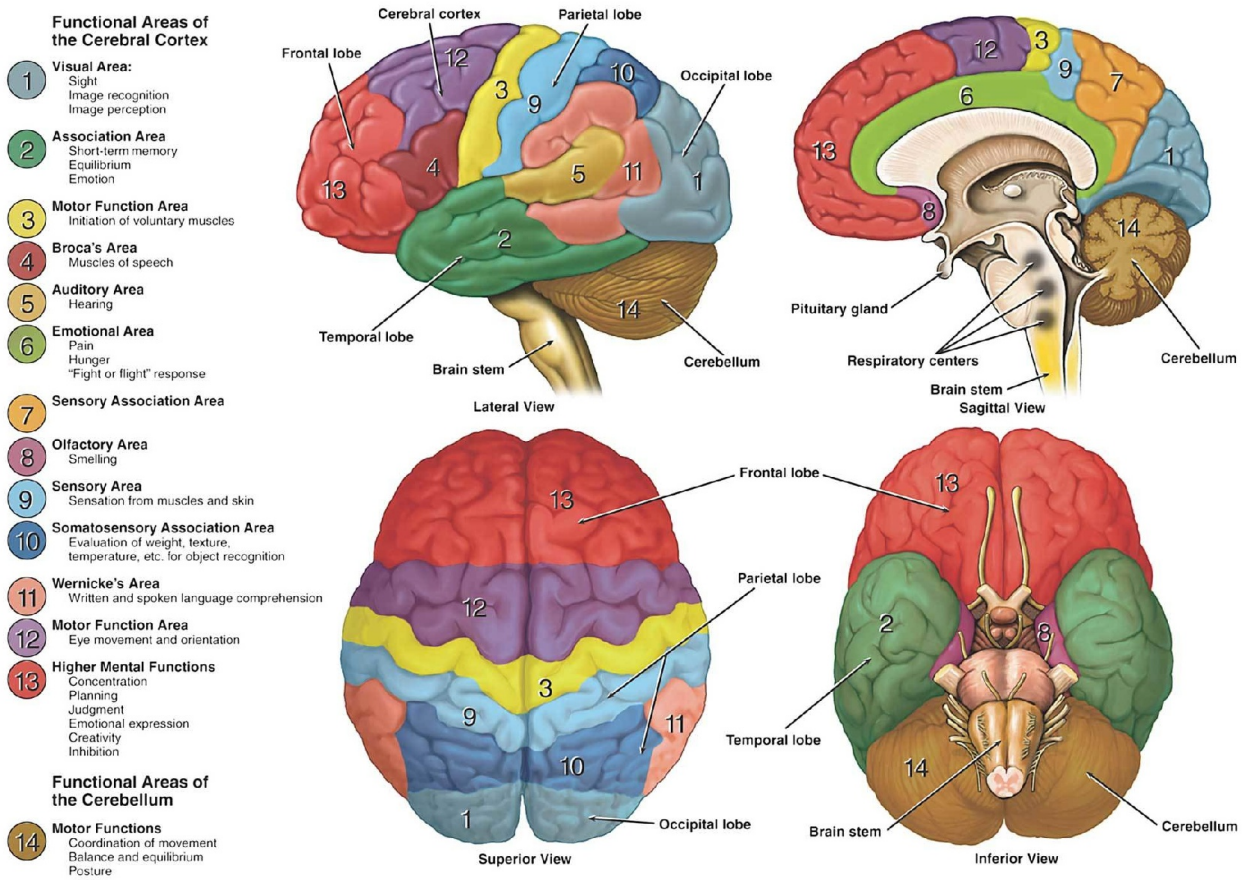


Figure 1.1: Human brain anatomy and functional areas of the cerebral cortex. Credits to [268]

Gray matter mainly contains neuronal cell bodies and relatively few myelinated axons (connecting brain regions), contrary to White Matter (WM). WM involves glial cells and myelinated axon fibers connecting the different regions of the brain, and playing support function to the neurons (e.g. by providing nutrients to the neurons). The brain also contains several deep structures associated with important cognitive tasks, in particular: the hypothalamus that regulates body temperature, synchronizes sleep patterns, and controls hunger and thirst; the amygdala that regulates emotion and memory and is associated with the brain's reward system and stress; hippocampus supporting memory, learning, navigation, and perception of space; ventricles filled with cerebrospinal fluid (CSF) that facilitates the transmission of several substances across brain areas.

What does structural MRI measure? This imaging technique uses the phenomenon of nuclear magnetic resonance (NMR) of the hydrogen atom to produce high-resolution, detailed images of internal brain structures and tissues. The strength of the magnetic field determines the resolution of the images. sMRI provides good contrast between gray matter and white matter. Nevertheless, it does not inform about white matter structure, which is measured by another modality (diffusion MRI). Concretely, sMRI gives in each voxel (3d volumetric

unit in a brain image) a tissue contrast that can be then pre-processed to derive measures of interest (such as gray matter density, cortical thickness or other surface-based measurements, see below).

1.1.2 Voxel-Based Morphometry

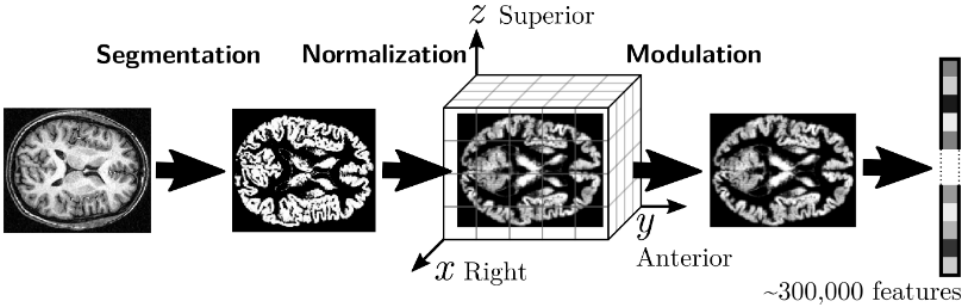


Figure 1.2: Features extraction with VBM pipeline.

This pre-processing has been described in [12], and it allows to extract a probability of tissues (gray matter, white matters) densities in each voxel from sMRI scan (see Fig. 1.2). It includes three main steps: **segmentation**, spatial **normalization** and **modulation**. The main idea is to extract gray/white matter tissue from sMRI and to apply a spatial deformation field to the image so that there is a spatial correspondence of voxels across subjects. VBM features are then aligned across subjects, and they can be used for downstream analysis (*e.g.,* statistical tests or machine learning with linear models, see section 1.2).

Segmentation consists in classifying each voxel according to the tissue it belongs to (gray matter, white matter, or cerebrospinal fluid). Spatial normalization is a composition of two transformations: i) a linear transformation that accounts for global alignment (rotation, translation, and global brain size); ii) a non-linear deformation that locally aligns brain structures (*e.g.,* DARTEL [12], HAMMER [255]). Note that step ii) expands and contracts locally brain regions. As a result, the normalized image needs to be scaled by the amount of contraction so that the total amount of GM is preserved. This final step is called modulation. In practice, it corresponds to multiply the normalized image by the Jacobian of the transformation. If the global brain size is not of interest (as it is the case in our experiments), one should apply a proportional scaling according to the individual Total Intracranial Volume (TIV), as post-processing, to fully modulated images.

Throughout this thesis, VBM pre-processing was performed with Computational Anatomy Toolbox (CAT [109]). This toolbox of Statistical Parametric Mapping (SPM) uses a modified segmentation procedure reducing the role of tissue priors. Although, it uses DARTEL for the normalization, CAT uses existing DARTEL templates in MNI space (as opposed to study-specific templates). This may seem somewhat sub-optimal, however, good performances have

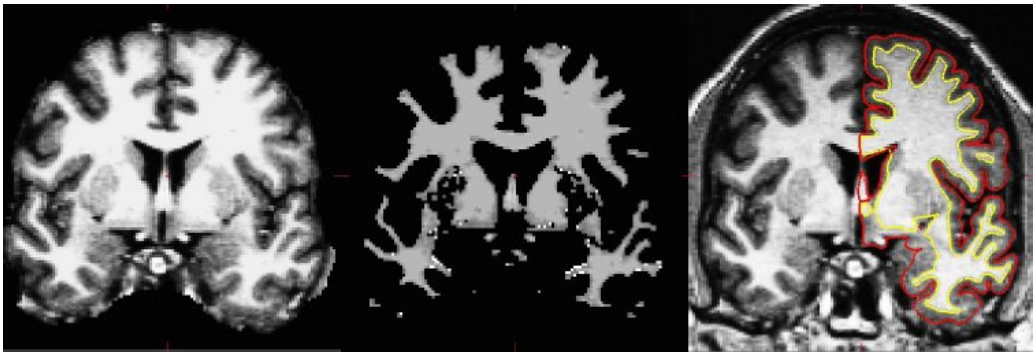


Figure 1.3: The three-stage from FreeSurfer cortical surface-based analysis. (Left) Skull-stripped image. (Middle) White matter segmentation. (Right) Surface between white and gray (yellow line, the white surface) and between gray and pial (red line, the pial suface) overlaid on the original volume. Once these two surfaces are reconstructed, surface-based measures can be computed on the FreeSurfer template (*e.g.,* cortical thickness, local curvature, surface area). Credits: Fischl and Dale [72, 97]

been reported [93] and the use of the same template for all studies offers the possibility to pool data across studies for subsequent statistical analysis.

Data quantity. VBM pre-processing produces hundreds of thousands of features (typically 300 000 GM voxels at $1.5mm^3$ resolution) representing the local GM volume at each voxel. Compared to the typical number of subjects in clinical datasets (rarely above 1k), we can easily anticipate a high risk of over-fitting with Machine Learning (ML) models, necessitating strong regularization and prior knowledge during training. We will come back to this issue in section 1.2. Nevertheless, producing VBM features require less *a priori* assumptions, *e.g.,* than region-of-interest (ROI) approach.

1.1.3 Cortical Surface-Based Morphometry

VBM features are based on tissues' concentrations and/or volumes, and they give only one piece of information about brain anatomy. Other surface-based cortical measures can be derived from sMRI scan. In particular, cortical thickness, surface area, or local curvature are also relevant anatomical features for brain imaging analysis. They characterize the amount of cortical atrophy or gyrification abnormality in brain regions that, ultimately can be useful to pinpoint underlying physiopathological processes in brain disorders (*e.g.,* schizophrenia or bipolar disorder). To derive such features from brain sMRI, FreeSurfer toolbox estimates two surfaces: the *white surface* that delimits gray matter from white matter (using sMRI contrast) and the *pial surface* that delimits gray matter from CSF (see Fig. 1.3) by nudging white surface. Once these two surfaces have been reconstructed from brain scan, surface-based features can be computed (*e.g.,* cortical thickness as distance between the two surfaces, local curvature of each surface, etc.).

The detailed pipeline is described in [72, 97]. All surface-based measurements maps are registered on the default template of Freesurfer. Thus, the dimensionality of the output features

is very high ($\approx$ 300 000), since it corresponds to the number of vertices on the cortical mesh of the brain. Consequently, the same over-fitting issue may appear for ML models (as previously with VBM features).

1.1.4 Does sMRI help to investigate brain disorders ?

We previously described the available features that structural MRI offers to conduct brain analysis. In this thesis, we focus particularly on subject-level prediction of brain disorders using sMRI so a natural question that arises is: do these anatomical features are related to brain disorders ? To answer to this question, we focus on two main brain disorders: **schizophrenia** and **bipolar disorder** (BD).

Findings for schizophrenia. Back to 1976, the first CT study of schizophrenia showed lateral ventricles enlargement in schizophrenia [162], confirmed later on with MRI. Global brain volume was also found significantly reduced compared to healthy controls. More fine-grained analysis using VBM and Regions-Of-Interest (ROI) revealed a decreased volume in frontal and temporal lobes [128, 146, 232, 256]. Sub-cortical structures such as amygdala and hippocampus were also reduced in schizophrenia patients. In a large meta-analysis conducted by [146], almost 50% of the studies involved revealed gray matter deficits in the left superior temporal, parahippocampal and inferior frontal gyrus. Abnormalities in the parietal and occipital lobes have also been reported but less consistently across studies. Finally, a more recent large-scale analysis [285] found large deficits in the volume of the hippocampus, amygdala, thalamus and accumbens in schizophrenia. They also discovered positive associations between increase of the volume of the putamen and pallidum volume in schizophrenia patients and duration of illness and age.

Findings for BD. Brain alterations have been consistently reported in sub-cortical structures such as hippocampus, thalamus and amygdala in subjects with BD compare to healthy controls [129, 130, 232]. In the largest study to date, ENIGMA consortium revealed that, on average, there is higher bilateral ventricular volumes and lower hippocampal, amygdala and thalamic volumes in BD vs HC [143]. Also, no structural brain differences were detected between BD sub-types (BD-I, BD-II and BD-NOS). As for cortical regions, lower cortical thickness in the anterior cingulate, para-cingulate, superior temporal gyrus and prefrontal regions were associated with BD [132, 220]. Again in a large-scale study, ENIGMA consortium confirmed previous findings concerning cortical thinning in frontal and temporal regions but also made new findings in inferior parietal, fusiform and inferior temporal regions. These regions are notably associated with disruption in sensorimotor integration, language and possibly emotion perception and rapid mood changes [58].

Overall, all these findings suggest that anatomical features are indeed well-suited to study brain disorders as it provides important information to pin-point discriminative brain regions

and possibly related them to functional analysis. However, it is important to note that previous observations were **valid at the group-level**, while we focus here on predictive models at the subject-level, a somewhat more difficult task.

1.2 Traditional machine learning

1.2.1 What is machine learning ?

Machine Learning (ML) is a sub-field of computer science whose goal is to learn from past experience in order to make predictions on future input. According to Tom M. Mitchell: "*a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E*". In this regard, we do not expect the machine to imitate or reproduce human intelligence but rather to make accurate predictions on a given task, without being explicitly programmed to do so. As a result, it is opposed to "standard" programming where a set of rules is explicitly written in order to compute a prediction from an input. In practice, it avoids writing a tremendously large program to perform a given task as it can automatically learn these rules (sometimes even in cases where they are unknown).

In ML, an algorithm is trained with data for a given task and we expect this algorithm to generalize well on new, unseen data, *i.e.,* to make accurate prediction for some task T on new data. Generalization is thus a fundamental concept in machine learning and it has been studied with statistical learning theory [288] by mainly relying on the complexity of the model[1] (*i.e.,* its Vapnik-Chervonenkis or VC-dimension).

In the following, we start by presenting traditional ML algorithms that fall under the supervised learning paradigm, that is, learning a target output from an input. They are widely used for brain imaging data analysis because of their simplicity (both in terms of interpretability and complexity). We then state the limit of such approaches and turn into more general representation learning models, in particular deep neural networks in the following section.

Supervised learning. Let $\{(x_i, y_i)\}_{i\in[1..N]}$ be a set of N labeled examples, *i.e.,* a set of annotated pairs where $x_i \in \mathcal{X}$ represent input data and $y_i \in \mathcal{Y}$ its corresponding annotation. We assume that these pairs are sampled from a joint distribution $p(X, Y)$ defined over $\mathcal{X} \times \mathcal{Y}$. In a classification problem, $(y_i)_{i\in[1..N]}$ are discrete (*i.e.,* $\mathcal{Y}$ is finite) while in a regression problem the labels $(y_i)_{i\in[1..N]}$ are continuous ($\mathcal{Y} = \mathbb{R}$). The goal is to learn a mapping from x to y such that future unseen input x' will be correctly mapped to its annotation y'. The natural questions are then: what model do we chose to map x to y ? How do we learn such mapping ?

[1] We shall remark here that this theory currently fails to explain the generalization capacity of state-of-the-art models and it is prone to intense debate in the community [317]

1.2.2 Linear models

Linear models learn a mapping $f_\theta(x) = \sum_{i=1}^{d} \theta_i x_i = \theta^T x$ that is a weighted combination of input data, assuming here $x \in \mathcal{X} \subset \mathbb{R}^d$. The learning rule is obtained by minimizing the (empirical) risk of the model f_θ on the training set $\mathcal{D} = \{(x_i, y_i)\}_{i \in [1..N]}$ through a loss function ℓ :

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i) \tag{1.1}$$

Here ℓ depends on the nature of the target y (continuous or discrete) but in all cases, it quantifies the error between the prediction made by the model $f_\theta(x)$ and the true label y.

Regression. If $y \in \mathbb{R}$ is continuous, then ℓ_2 squared loss $\ell(f_\theta(x_i), y_i) = (f_\theta(x_i) - y_i)^2$ leads to a convex objective. It is known as the Ordinary Least Square (OLS) regression and has the following solution: $\theta^* = \arg\min \mathcal{L}(\theta) = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ where $\mathbf{x} = (x_1, ..., x_N)^T$ and $\mathbf{y} = (y_1, ..., y_N)^T$.

Classification. If y is discrete, we assume it only has a binary value, 0 or 1 (extension to multi-class is not treated for simplicity). Two main losses can be used:

(i) the logistic loss, based on the probabilistic model $p_\theta(y = 1|x) = \frac{1}{1+\exp(-f_\theta(x))} = \sigma(f_\theta(x))$ where σ is the Sigmoid function. It can be expressed as the negative log-likelihood $\ell(f_\theta(x_i), y_i) = -\log p_\theta(y_i|x_i)$.

(ii) the Hinge loss, based on margin loss $\ell(f_\theta(x_i), y_i) = \max(0, 1 - y_i f_\theta(x_i))$ (here assuming that $y_i \in \{-1, 1\}$). It is notably used for training SVM (see next section).

Both logistic and Hinge loss are convex so any standard convex optimizer can be used.

Regularization technique as inductive bias

In a practical scenario with brain imaging, the number of training samples N is very small compared to input dimension d, *e.g.,* $N = 1000$ subjects vs $d > 30000$ voxels. In this scenario, the model f_θ also contains much more parameters than the number of observations since $\theta \in \mathbb{R}^d$. According to statistical learning theory [288], the generalization capacity of a model depends directly on its VC-dimension (*i.e.,* its complexity defined as the cardinality of the largest set of points that the algorithm can label arbitrarily) and N. For d dimensions, the linear model f_θ has a VC dimension $d + 1 \gg N$ so there is no good guarantees for generalization to new data. In other words, the model can perfectly fit the training data (*e.g.,* with 100% accuracy) but it may have random performance on new data, an issue also known as **over-fitting**.

A standard approach for fighting over-fitting is by imposing a penalty on the weights θ that depends on the prior we have about the final solution. This penalty $\mathcal{R}(\theta)$ is added to the loss

function ℓ so that the empirical risk becomes:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i) + \lambda\mathcal{R}(\theta) \tag{1.2}$$

with λ an empirical trade-off (also viewed as a Lagrangian multiplier when imposing $\mathcal{R}(\theta) < cst$) that needs to be set.

Ridge regularization. It imposes a ℓ_2 Euclidean squared penalty on the weights $\mathcal{R}(\theta) = ||\theta||_2^2$. It prevents to have exploding weights in the final solution that may have over-fitted on noisy voxels.

Lasso regularization. It imposes sparse solution through a ℓ_1 penalty on the weigts $\mathcal{R}(\theta) = ||\theta||_1$. For neuroimaging data, it is particularly suited when we expect only a few voxels to be predictive of a clinical outcome (*e.g.,* diagnosis). However, there is no spatial constraints on non-zeros weights.

ElasticNet. It tries to take the best of both (previous) worlds by imposing ℓ_2 Ridge and ℓ_1 Lasso constraints: $\mathcal{R}(\theta) = ||\theta||_2^2 + \lambda_1||\theta||_1$.

Total Variation. This regularization is widely used in image denoising and restoration. It accounts for the spatial structure of images by encoding piecewise smoothness and enabling the recovery of homogeneous regions separated by sharp boundaries. It is expressed as $\mathcal{R}(\theta) = ||\nabla\theta||_{2,1}$.

Interpretability

Linear models are appealing for their simplicity. They produce spatial weighted maps (through θ) that can be interpreted as patterns of activation (a.k.a. **predictive signature**), for instance for a binary classification task such as patients vs healthy controls. Nevertheless, raw predictive map of coefficients makes the interpretability challenging. Indeed, the magnitude of coefficients is difficult to interpret since it depends on many factors: the regularization, the size of the regions, etc. Moreover, some coefficients may be large but highly unstable across training with different subset of samples (*i.e.,* folds). Therefore, z-score map is often required to compute predictive coefficients to bypass the problem of magnitude and highlight the only most stable voxels (and regions).

1.2.3 Kernel-based models and application to Support Vector Machines

Kernel method for SVM

The previous linear models have an important limitation: they only model linear relationships between input data (*e.g.,* MRI voxels) to predict the target. As a result, features often need to

be hand-crafted from raw data in order to obtain good linear predictors of the target (which can be difficult to obtain, especially in our context with very high inter-individual heterogeneity for brain disorders and limited knowledge about the biomarkers involved). A first alternative was presented in 1995 by Cortes and Vapnik [70] where input data are projected to a very high (potentially infinite) dimensional feature space through a feature mapping $\phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ where $d' \gg d$. Instead of predicting a target y through linear combination of raw input data x, it is now predicted with the features $\phi(x)$. It is a first step towards working with raw data. The previous decision function f_θ can be written as[2]:

$$f_\theta(x) = \theta^T \phi(x) \tag{1.3}$$

It now becomes non-linear w.r.t input x but the optimization of the learning rule $\mathcal{L}(\theta)$ remains convex with convex loss function (i.e. we can still find a global solution). However, the question is: how do we define such mapping ϕ ? Cortes and Vapnik proposed to use a particular form of the loss function ℓ in order to indirectly define ϕ with a kernel function living in a Reproducible Kernel Hilbert Space (RHKS) space, thus introducing Kernel Support Vector Machines (Kernel SVM). They notably demonstrate that minimizers of the Hinge loss (margin-based loss function for classification problems) with ℓ_2 penalty on θ leads to a solution of the form:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d'}} \mathcal{L}(\theta) = \sum_{i=1}^{N} y_i \alpha_i \phi(x_i) \tag{1.4}$$

Where $\alpha_i \geq 0$ are parameters to find. This allow to re-write the decision function depending on ϕ *only* through dot-products:

$$f_{\theta^*}(x) = \sum_{i=1}^{N} y_i \alpha_i \phi(x_i)^T \phi(x) \tag{1.5}$$

We know that dot-product $\phi(\cdot)^T \phi(\cdot)$ defines a kernel $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ in a RKHS space. Reversely, Mercer's theorem ensures that any continuous symmetric non-negative definite kernel induces a dot-product in feature space. As a result, we can write the previous decision function using *only* a kernel K full-filling Mercer's condition:

$$f_{\theta^*}(x) = \sum_{i=1}^{N} y_i \alpha_i K(x_i, x) \tag{1.6}$$

This observation has an important practical consequence: we do not need to define explicitly features map ϕ but only a Mercer Kernel K, which is easier to craft. Multiple kernels have been designed over the years (e.g. Gaussian and polynomial are the two most famous) and they reflect the prior we have on input data. Intuitively, it defines a notion of similarity between pairs of data. This model is known as Kernel SVM and it learns a non-linear decision boundary for

[2]We omit the bias for simplicity since it does not change the reasoning.

classification problems. It has also been extended to regression by modifying the loss function accordingly (we refer to [14] for more details).

One important bottleneck for Kernel-SVM is that it does not scale well for large datasets. Its temporal and spatial complexity scale as $O(N^3)$ and $O(N^2)$ respectively (in particular for computing and storing the kernel matrix $\mathbf{K}_N = (K(x_i, x_j))_{i,j\in[1..N]}$). It can be prohibitively expensive when $N > 10^4$, which can be the case also for neuroimaging data (as we shall see in this thesis).

Generalization to other models. The main "trick" in Kernel SVM to go from a linear to a non-linear decision function is to 1) map input data in a high-dimensional space with feature map ϕ and 2) view the dot-product between features maps as the application of a kernel K, thus avoiding an explicit definition of ϕ. These 2 ingredients can be inserted in any ML algorithms involving dot-products between input data. A famous example is Principal Component Analysis (PCA). It is an unsupervised algorithm (*i.e.,* it does not require labels to learn) that decomposes input data on axis of maximal variance. It mainly operates by diagonalizing the covariance matrix $C = \sum_{i=1}^{N} x_i x_i^T$. Once again, all dot-products can be replaced by application of a kernel, giving rise to Kernel PCA.

In summary, we saw that traditional ML algorithms enjoy important desirable properties: linear models are interpretable, with strong theoretical guarantees and allow fast computations; kernel-SVM is a non-linear model with also strong convergence properties and versatile as to which kernels we can choose. In the next section we introduce deep neural networks as a broader class of algorithms, capable of modelling any bounded continuous decision function and performing feature extraction for a very wide range of tasks (both for unsupervised and supervised learning).

1.3 Deep representation learning

Deep learning or deep representation learning is a subfield of machine learning that gained tremendous attention in the last decade. As opposed to previous traditional machine learning algorithms, deep learning models learn a representation of raw input data to perform its task (such as classification, regression, clustering, etc.), thus mapping input data to a *latent space* with desirable properties (*e.g.,* linear separability of input data into classes for supervised classification). This mapping is learned in a layer-wise manner as we shall see, from low to high-level abstraction. One important implication is that deep learning models do not require human-generated features crafted from raw data to learn (as it was previously implicitly the case with linear models and, to a lesser extent, with kernel-SVM). For instance, in the context of neuroimaging, deep learning models would not require a computationally extensive pre-processing based on prior knowledge (e.g. anatomical knowledge through atlases in neuroimaging or non-linear registration to a template). This question will be studied in the first

chapter.

Deep learning algorithms have a long history that dates back to 1943 with McCulloch and Pitts [200] when they formalized the brain computation of a single biological neuron, firing when its weighted input signal is above a given threshold. A supervised algorithm with a learning rule was then invented by Rosenblatt in 1957 based on this model (the Perceptron [238]). The next three decades (until the 80's) allowed the development of the back-propagation algorithm (1960, by Kelley [168] used currently to train neural networks), the Convolutional Neural Network (1980, by Fukushima [104]) and its training with back-propagation algorithm (1989 by LeCun [188]).

The real breakthrough happened later, in 2012, when a deep learning algorithm won, by a large margin, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) for image classification. This challenge was launched two years before, in 2010, by Fei-Fei Li. It is based on the large-scale dataset ImageNet [74] that contained 3.2 million natural images at that time. This breakthrough was allowed thanks to three crucial factors: **computational resource availability** (in particular training on Graphics Processing Unit or GPU, invented in 2005 [266]), **data availability** (with ImageNet), and **model size** (with AlexNet [181], CNN with five convolutional layers). These 3 ingredients are the cornerstone of current performance and an improvement in each one of them leads to drastic increase in accuracy. As an illustration, ImageNet has grown in size from 3 to 14 million images, biggest models contain now more than 1 billion parameters and use hundreds of GPUs [117] for training. The accuracy on ImageNet increased from 63% (with AlexNet) to 91% (with Transformers).

1.3.1 Multi-Layer Perceptron

The Perceptron (invented by Rosenblatt) is a simple linear model with a Heaviside activation function at the end to make a binary prediction (0 or 1) from input (see section 1.2.2). It is biologically inspired by the functioning of a neuron in the brain. It can be written as $f_\theta(x) = \phi(w^T x + b)$ where $\phi(x) = 1$ if $x > 0$ and 0 otherwise (Heaviside activation function). $\theta = \{w, b\}$ are the parameters to learn and f_θ is the decision rule.

2-layers Perceptron. Previous model is simple and outputs only a single value. The main innovation comes when we compose 2 Perceptrons $f_\theta = f_{\theta_1} \circ f_{\theta_2}$ with $\theta = \{\theta_i\}$ the parameters to learn. In this case, each Perceptron f_{θ_i} can output multiple values[3] $f_{\theta_i}(x) = \phi(W_i x + \mathbf{b}_i)$ where $W_i \in \mathbb{R}^{d_{i-1} \times d_i}$ is a matrix and $\mathbf{b}_i \in \mathbb{R}^{d_i}$ a vector. We noted d_0 the input dimension, d_1 the *hidden layer dimension* and d_2 the output dimension. In that case, the model has d_1 neurons in its intermediate hidden layer (see Fig. 1.4).

Going from one to two layers is a crucial conceptual and mathematical shift from traditional machine learning algorithms. In short, it allows to learn a *representation* of the data to perform a prediction task. Indeed, by learning jointly $\{\theta_1, \theta_2\}$, the model learns to map x to a latent

[3] We use a slight abuse of notations as ϕ is now applied point-wise on a vector.

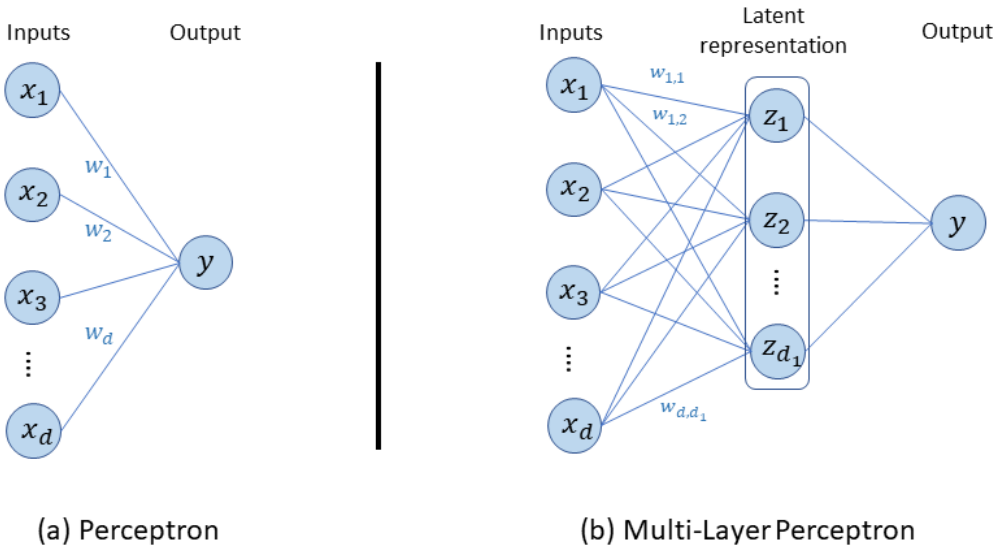


Figure 1.4: From Perceptron to Multi-Layer Perceptron: a latent representation is born. Universal Theorem ensures that 2-layers Perceptron can model any continuous function on a bounded space: it allows much more representation capacity than shallow Perceptron.

space through $f_{\theta_1}(x)$ before actually predicting the scores with f_{θ_2}. For instance, for a supervised classification task, f_{θ_1} should output a representation of input data linearly separable (by f_{θ_2}) for each class.

Universal Approximation Theorem. [71] Mathematically, 2-layers Perceptron enjoys the Universal Approximation Theorem stating that the decision function f_θ can approximate *any arbitrary* continuous function on a bounded space, given a sufficiently large width (*i.e.,* high hidden dimension d_1). It is true for any non-constant bounded activation function ϕ.

This theorem is very general and it states the existence of an optimal 2-layers Perceptron for a very large set of tasks but it does not specify how to build it (architecture, number of neurons, etc.). As we shall see, the representation capacity of such model comes at a cost: the risk of over-fitting (see section 1.2.2). Briefly, for big enough model (large width), it can perfectly learn all data in the training set by memorizing it (based on spurious features), with very poor generalization performance on new incoming data. More broadly, even current deep models (with hundreds of hidden layers and state-of-the-art architectures) could theoretically easily over-fit and understanding their impressive generalization performance is still an open problem [317] ("*understanding deep learning requires rethinking generalization*").

Multi-Layer Perceptron. The Multi-Layer Perceptron (MLP) generalizes the previous idea to an arbitrary number of layers. We can compose k Perceptrons together such that $f_\theta = f_{\theta_1} \circ f_{\theta_2} \circ ... \circ f_{\theta_k}$ with $f_{\theta_i}(x) = \phi(W_i x + \mathbf{b}_i)$ as before. In that case, multiple intermediate representations are defined after each hidden layer until the last layer where the actual task is performed. A naive question would be: why do we care about stacking multiple layers if

2-layers Perceptron has enough representation capacity ? The answer is mostly empirical as mathematical analysis of MLP is often limited to 2 or 3 layers [6]. Indeed, the past decade of research has shown that "*layerwise stacking of feature extraction often yielded better representations, e.g., in terms of classification error [90, 185], quality of the samples generated by a probabilistic model [239] or in terms of the invariance properties of the learned features*" [30]. It is also built on the prior that *"concepts that are useful for describing the world around us can be defined in terms of other concepts, in a hierarchy, with more abstract concepts higher in the hierarchy, defined in terms of less abstract ones"* (as observed by Bengio in 2013 [30]).

In practice, the idea behind MLP remained (that is: building a layer-wise representation of input data for achieving a given task) but its actual implementation with current deep neural architectures has largely evolved over the years. It has been driven by empirical observations, intuitions coming from cognitive science or biological systems, and engineering tricks to arrive at the current architectural choice.

Deep neural networks optimization. How do we train such multi-layers model ? As for traditional ML algorithms, a loss function ℓ needs to be defined such that we minimize the empirical risk $\mathcal{L}(\theta)$ (see section 1.2.2 for a definition in the supervised context). Nevertheless, as the reader may have noticed, MLP is a highly non-linear and non-convex model (*e.g.,* since it can theoretically represent any continuous function on a bounded space). As a result, the search for global minima, if they exist, is difficult and often impossible without any further assumption on the architecture. Instead of looking for global minima, the intuition is that prior knowledge implemented through deep architecture allows to define a starting point (in parameters' space) in the bassin of attraction of "good" local minima, where "good" means low generalization error. Keeping in mind such intuition, the optimization procedure is Stochastic Gradient-Descent (SGD) [233] and the update of the weights follows the rule:

$$\theta \leftarrow \theta - \alpha \nabla \mathcal{L}(\theta) \tag{1.7}$$

where α is called the learning rate. One important advantage of SGD is scalability: it allows to learn from a very large-scale dataset by decomposing the data into several chunks or "batch" and to approximate the gradient $\nabla \mathcal{L}(\theta)$ using such batch of data (and not the entire dataset). Nevertheless, its main drawback is the differentiability assumption: it supposes that the model f_θ is differentiable almost everywhere (*i.e.,* we can compute its gradient w.r.t θ). In particular, it limits the architecture of the deep neural networks (*e.g.,* the activation function ϕ in the previous MLP model). Originally, ϕ was defined as Heaviside step function but later on, ReLU function [114] (that fires only if input is positive, like Heaviside, but proportionally to the input) was introduced to impose sparsity inside representation, a hypothesis more biologically plausible compared to previous sigmoid and hyperbolic tangent activation functions. Empirically, it led to very good performance and it is still used in modern architecture even if variations have

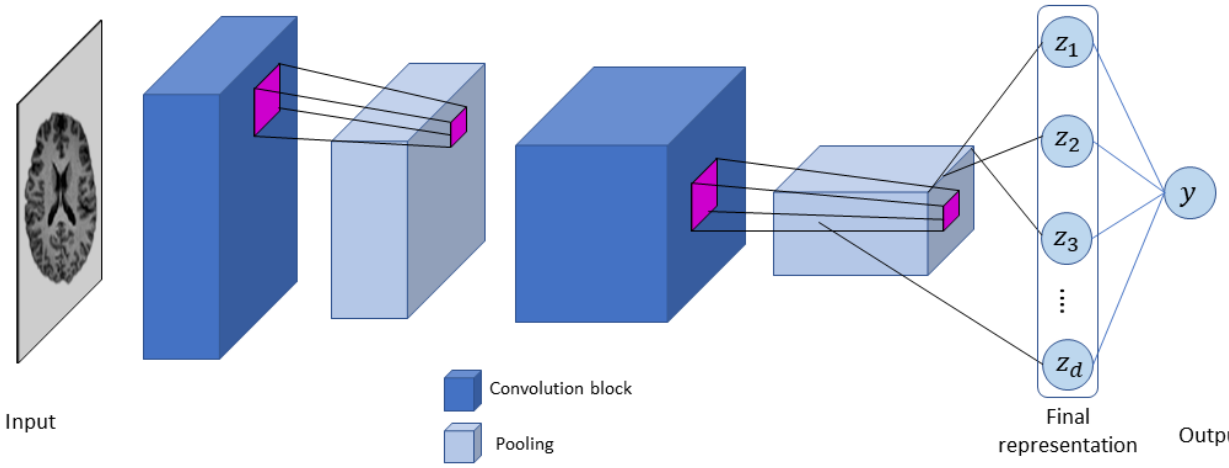


Figure 1.5: Illustration of a CNN integrating several blocks of convolution layer followed by activation function and pooling. Like for 2-layers Perceptron, the final representation is mapped to the output with a fully-connected layer.

been proposed over the years (*e.g.,* LeakyReLU [195], GELU [140], etc.).

We previously saw that MLP is an attractive model as it allows to perform deep representation learning in a layer-wise manner, enjoying an exceptional representation capacity of a large class of functions. In what follows, we present a very successful sub-family of models, introduced very early on in 1980 by Fukushima [104] and then developed by LeCun in 1989 [188]: Convolutional Neural Networks (CNN).

1.3.2 Convolutional Neural Networks

CNN is a sub-family of MLP that takes inspiration from biological cortical neurons inside the visual cortex of animals [104, 151]. A neuron responds to a stimuli located in a very restricted region in the brain including only few neurons, known as its receptive field. If we transpose this observation to MLP, it means that an artificial neuron needs not to be connected to all neurons in the previous layer but rather to a few adjacent ones, defining its own artificial receptive field. Mathematically, it corresponds to re-write the matrix-vector multiplication $W_k x$ in the k-th layer $f_{\theta_k} = \phi(W_k x + \mathbf{b})$ by a **convolution** operation $K * x$ where K is now called a kernel and has a much lower size than the original matrix $W_k \in \mathbb{R}^{d_{k-1} \times d_k}$ (following previous notations). Convolution is a mathematical tool often used in signal processing (in particular for filtering) as it has the elegant property of transforming point-wise multiplication in frequency domain (*i.e.,* after Fourier transform) to convolution in time domain, known as the convolution theorem.

Example. If x is a 2D image represented as a matrix $x \in \mathbb{R}^{H \times W \times C}$ of height H, width W with C channels (*e.g.,* Red, Green, Blue for natural images), then we can define a kernel $K \in \mathbb{R}^{h_K \times w_K \times C}$ where $h_K \ll H$ and $w_K \ll W$ set the receptive field of all neurons for the k-th

layer. In a standard MLP, $h_K = H$ and $w_K = W$, and the number of parameters to train in the k-th layer is $H \times W \times C$ which is the input size. The convolution operation is defined as:

$$(K * x)[i, j] = \sum_{c=1}^{C} \sum_{r=1}^{h_K} \sum_{s=1}^{w_K} K[r, s, c]x[i - r, j - r, c] \tag{1.8}$$

which is well-defined for all $i, j \in [1..H] \times [1..W]$ if we add zero-padding around image x. The key point here is that the receptive field $h_K \times w_K$ is very small compared to the entire image size (typically 3×3 or 7×7 for input image of size 32×32 or 256×256). As a result, kernel size $w_K \times h_K \times C$ contains far less parameters to learn and the whole network architecture is much lighter than its fully-connected MLP counter-part, for the same number of layers.

In the previous example, we defined only one kernel to output a *features map* from an input image (with the same size as input if we add zero-padding). Each feature in this features map is a local aggregation of input pixels (for 2D image or voxel for 3D image). We can generalize this idea to multiple kernels in order to output several features maps. A convolution layer with C' kernels then outputs C' features maps from input image x that, when concatenated on the last dimension, gives a tensor of size $H \times W \times C'$.

Sparsity. Re-writing the matrix-vector multiplication $W_k x$ by $K * x$ notably implies strong sparsity in the neural connections. Indeed, since convolution is a linear operator, we can always see it as a matrix-vector multiplication with a very sparse circulant matrix. In fact, as we saw in the last example, a neuron in each layer is connected to a very small subset of neurons in previous layer (belonging to its receptive field). Consequently, this model removes most of the connections in standard MLP by including prior knowledge on spatial arrangement of the neurons. Another consequence of convolution is **weights sharing**: the same kernel is used to compute all features in the features map (which is another way of seeing sparsity).

Pooling. Another key ingredient is missing to define the building block of modern CNN architectures: pooling operation. It allows more spatial invariance by reducing the resolution of features map using mainly averaging or max-pooling [181, 245] over the neighborhood of each feature in the features map. This operation is performed after the activation function such that the k-th layer is $f_{\theta_k}(x) = \beta(\phi(K * x + \mathbf{b}))$ where $\beta(\cdot)$ is a pooling operation that down-scales the features map.

Equivariance and invariance to translation. CNN has two useful properties, intrinsic to its architecture: it is equivariant to any translation and also "mostly" invariant to small translations. Equivariance means that, for any translation T of an input x, the model f_θ fullfills $T(f_\theta(x)) = f_\theta(T(x))$. It is true since all convolution layers are equivariant to T. Invariance is more subtle and comes from pooling. As we aggregate close features in a features map with pooling, changing input x with a small translation T will likely not change the pooled

values. This property still highly depends on pooling size and input so it is not as general as equivariance for CNN.

Modern architectures

The basic CNN architecture depicted in Fig. 1.5 is the simplest one, combining only convolution layer with activation function and max-pooling (architecture used in AlexNet in 2012 to win the ILSVR challenge). Several tricks led to major improvement in performance during the next years. We rapidly expose one of the main modern architectures but we refer the interested reader to a recent review [8] for an in-depth analysis. These architectures will be notably compared in the first chapter on brain imaging data.

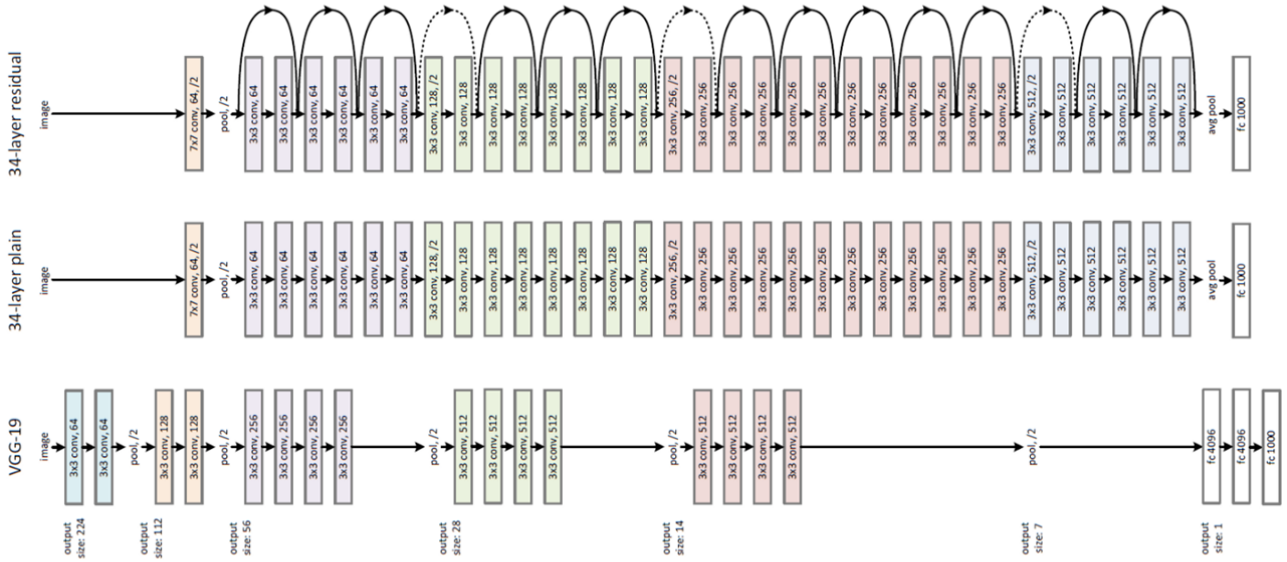


Figure 1.6: VGG uses small kernel size to build deeper model and achieve better performance on ImageNet. ResNet introduces skip-connection between convolution block as a novel way to avoid vanishing gradient during training. It allows to train very deep networks with more parameters, while still achieving better generalization than VGG. Credits to [135]

Visual Geometry Group (VGG) [258] This network was introduced in 2014 by Simonyan and Zisserman and they essentially demonstrate two main properties in CNN: i) increasing depth (*i.e.,* by stacking more convolutional layers) helps to generalize better and ii) use of small kernel size improves performance. In practice, they used up to 19 weight layers to hold the first and second place in ILSVR-2014 Challenge on localization/classification task and 3×3 kernels inside convolution layers. This is the smallest receptive field possible to "*capture the notion of left/right, up/down, center*" [258]. Using smaller kernel size than in previous networks

(*e.g.,* AlexNet [181] with 5×5 or 11×11 in early layers) allows to use deeper networks for the same number of parameters.

ResNet [135] The quest for deeper networks to achieve better generalization performance encountered an important optimization issue: *vanishing gradient* [29, 113]. During training, the gradients associated to early layers weights is smaller and smaller as the depth increases, leading to poorer performance for very deep CNN trained with back-propagation algorithm since first layers weights barely change during optimization. Bengio hypothesized [30] that this issue "*is centered on the singular values of the Jacobian matrix associated with the transformation from the features at one level into the features at the next level [113]. If these singular values are all small (less than 1), then the mapping is contractive in every direction and gradients would vanish when propagated backwards through many layers*". A simple, yet effective idea introduced by He et al. [135] solved this issue: for each convolution block, they added identity mapping between input and output of this block (a.k.a residual skip-connection). Mathematically, it consists in re-writing the k-th layer as[4] $\tilde{f}_{\theta_k}(x) = f_{\theta_k}(x) + x$. This way, the gradients can "flow" backwards until the very first layers, for an arbitrary depth size. In particular, He et al. tested until 152 layers (ResNet152) which hold the best results on ImageNet classification task.

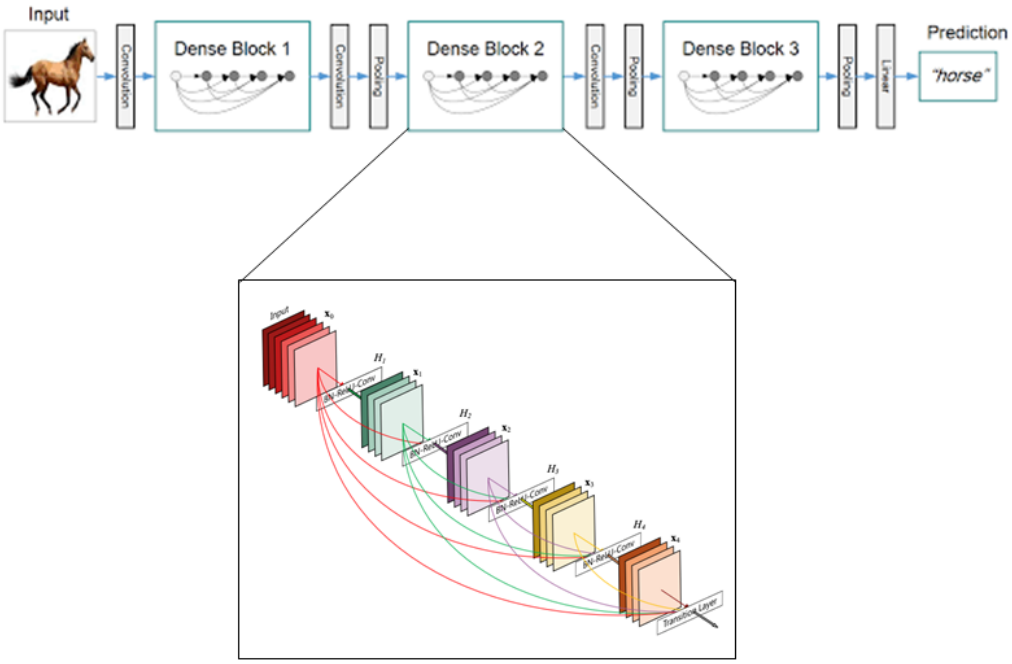


Figure 1.7: DenseNet architecture. Credits to [149]

DenseNet [149] ResNet solved the vanishing gradient issue and it allowed the training of very deep architectures. The main shortcoming is that simply stacking more layers add parameters and it may lead to over-fit at some point. Huang et al. [149] proposed features re-using as a new

[4] We hypothesized here that $f_{\theta_k}(x)$ has same dimension as input x. It is generally true for convolution layers if input/output channels size match.

way to build more compact convolution blocks, taking benefit of the representation capacity of smaller-size CNN to increase generalization capacity with less layers (and parameters) than ResNet. In essence, the main idea is to concatenate *all* past features maps (with same size) inside a convolution block and to apply convolution layer to this concatenated representation. As a result, we hope that all past relevant features already learned during training will be re-used (and not redundantly learned) in the next layers.

1.3.3 Self-supervised learning

Limits of supervised learning. In previous section 1.2.2, we presented a learning rule for supervised problems, when annotations y are available for all input data. In that case, the loss function $\ell(f_\theta(x), y)$ gives the error between prediction $f_\theta(x)$ and true annotation y. In deep representation learning, it means we want the penultimate layer to output a data representation as much predictive as possible of y, in a linear manner. This approach has 2 main limitations: it requires massive amount of data to converge towards a "good" (*i.e.,* generalizable) solution; the learned representation is only adapted to one task, and may not be suited to other "related" ones ("related" needs to be defined). The first point is critical especially in the medical domain where large annotated datasets are rare and costly (*e.g.,* brain MRI of patients with brain disorders in our context). The second point concerns a shift in paradigm where we do not seek to find a representation only predictive of a single supervised signal y, but rather one that can be applied to many different tasks (potentially in different input domains). We emphasize that, for natural images, strong correlations have been found [177] between supervised pre-training accuracy on ImageNet and several downstream classification tasks performance on new datasets. Nevertheless, we argue that such findings i) are limited to natural images and may not yield on medical images [230] and ii) may not be optimal for distinct tasks (e.g. object detection or semantic segmentation [91]). More concretely, we will check that ImageNet pre-training is not adapted in our context in Chapter 3 (corroborating our previous hypothesis).

What is self-supervised learning ? As Y. LeCun stated in its recent "path towards autonomous machine intelligence" [187], "*self-supervised learning is a paradigm in which a learning system is trained to capture the mutual dependencies between its inputs. Concretely, this often comes down to training a system to tell us if various parts of its input are consistent with each other.*" As a result, it is an unsupervised approach (*i.e.,* it does not require human annotations to learn) that learns a data representation "relevant", *i.e.,* generalizable to a large set of downstream tasks, hopefully on multiple domains. It tries to solve the two main issues of supervised learning previously discussed.

Self-supervised models requires two ingredients that need to be set: the observed part x of an input (can be image, text, audio, etc.) and another–possibly unobserved– part y[5]. An

[5]We use the same notation y as before in the supervised context on purpose: here it can be considered as an "artificial" label we aim to retrieve.

important remark is that the model is not expected to *predict* y from x but rather to tell us the degree of *compatibility* between x and y, as y may be only one answer among an infinite number of plausible ones. Building pairs (x, y) give the "pretext task" the machine is expected to solve like in supervised learning.

To give more concrete examples of self-supervised models, we divide them into two main categories specially dedicated to visual representation learning (borrowed from the complete survey by Jing et al. [160]) :

- **Context-Based methods.** x is a part of an input image that either i) share the same visual context than an other part y (context similarity algorithms) or ii) is predictive of spatial context information y (spatial context algorithms);
- **Generation-Based methods.** x is an input image (or a sub-part) and y is the original image. These algorithms thus learn to generate image y from x.

Context-based methods. Popular **context-similarity models** define several groups of data that share the same semantic features and are trained to map each group of data to the same region in CNN latent space. Such groups can be defined for example with a strong data augmentation strategy (*e.g.,* crop or color distorsion) such that each group contain only augmented versions of the same original image. Pairs (x, y) are then defined as all possible pairs of data inside the same group. The way this mapping is learned can be with clustering algorithm (SwAV [43] or Deep Cluster [42]), cross-entropy (SimCLR [52] or MoCo [136]), Euclidean distance (BYOL [120]), variance reduction (VICReg [22] or Barlow Twins [314]). During training, the CNN must be invariant to the class of samples belonging to the same group, thus implicitly learning semantic information about images. For self-supervised models using **spatial context** of an image, the pretext task usually consists in predicting the relative position of two random patches x and y inside the same image [83]. More complicated puzzles have been proposed, such as solving the Jigsaw puzzle [208] but they are based on the same original idea. Pretext task with the full image x can also be crafted, *e.g.,* by rotating x of an angle y and learning this angle [111].

Generation-based methods. The simplest generation-based model is the auto-encoder [145]. Input x is the same as output y (an original image) and the task consists in compressing the data by encoding x with a CNN to produce a small latent code, then decoded to generate y (the original input). This latent code has a very small size compared to input x and we expect it to contain semantic information. Other models have been proposed later on, whose main idea is to degrade an original image y to produce an image x that should be predictive enough to retrieve the original y. For instance, inpainting [216] consists in retrieving missing regions inside an image. These regions are randomly removed from y using black squares for instance (a.k.a. cutout [76]). The machine is expected to learn the color and structure of common

objects inside images to perform the task. Colorization [186] is based on a similar idea: the task is to predict pixels color from a gray-scale image x, based on its semantic. It thus requires the recognition of objects and semantic regions clustered together that have the same color. These methods also rely on encoder-decoder architecture to perform the pretext task.

Self-supervised learning for medical imaging. Multiple pretext tasks have been crafted specially dedicated to medical imaging. They can take advantage of 3D image spatial structure to define context-based methods such as playing the Rubik's cube [272, 321] (*i.e.,* decomposing input image into sub-volumes randomly shuffled and learn to reassemble them) or new context-similarity models leveraging local regions inside input images to define semantically similar groups [47] (particularly useful for brain segmentation tasks). Generation-based models have also been proposed for medical imaging [50, 320] where the main innovation comes from the design of transformations applied to original image y, in order to produce degraded version x. As before, the model is trained to predict y from x. For instance, Zhou et al. [320] proposed non-linear transformations, pixel shuffling and cutouts to learn respectively appearance, textures and context from both segmentation and classification downstream tasks. An extensive comparison between these models for brain segmentation and diabetic retinopathy detection has been presented by Taleb et al. [270].

1.3.4 Transfer learning

As we saw in the Introduction, the main goal of this thesis is to learn a (deep) representation of brain imaging data of the healthy population in order to better discriminate patients with brain disorders from healthy controls. This paradigm follows the general principle of Transfer Learning [28, 44, 310] where one seeks to pre-train a deep model on a source domain $\mathcal{D}_S$ with a source task $\mathcal{T}_S$ in order to improve the final representation on the target domain of interest $\mathcal{D}_T$ and a target task $\mathcal{T}_T$. The main assumption in TL is that $\mathcal{D}_T \neq \mathcal{D}_S$ or $\mathcal{T}_T \neq \mathcal{T}_S$ (if both are equals, it would fall into the traditional machine learning setting). TL is initially inspired by Multi-Task Learning [44] (MTL) where source and target domains are equal $\mathcal{D}_T = \mathcal{D}_S$ and the model is trained on multiple tasks simultaneously using the *same representation* for all tasks (see Fig. 1.8). The main assumption is that common features should be extracted to perform correctly the tasks so the model can exploit common statistical properties between tasks to improve the final representation.

TL is somewhat more general than MTL as source and target domains can be different, but the underlying assumption is similar to MTL. Several categories exist for TL, depending on whether $\mathcal{T}_S = \mathcal{T}_T$ (thus $\mathcal{D}_S \neq \mathcal{D}_T$, called "transductive transfer learning" and it can be related to domain adaptation [214]) or $\mathcal{T}_S \neq \mathcal{T}_T$ (called "inductive transfer learning"). In our case, we clearly fall under inductive transfer learning since we do not assume to have access to patients with brain disorders during pre-training. As suggested in the previous section, our main approach will use self-supervised learning for pre-training on source domain $\mathcal{D}_S$. The main

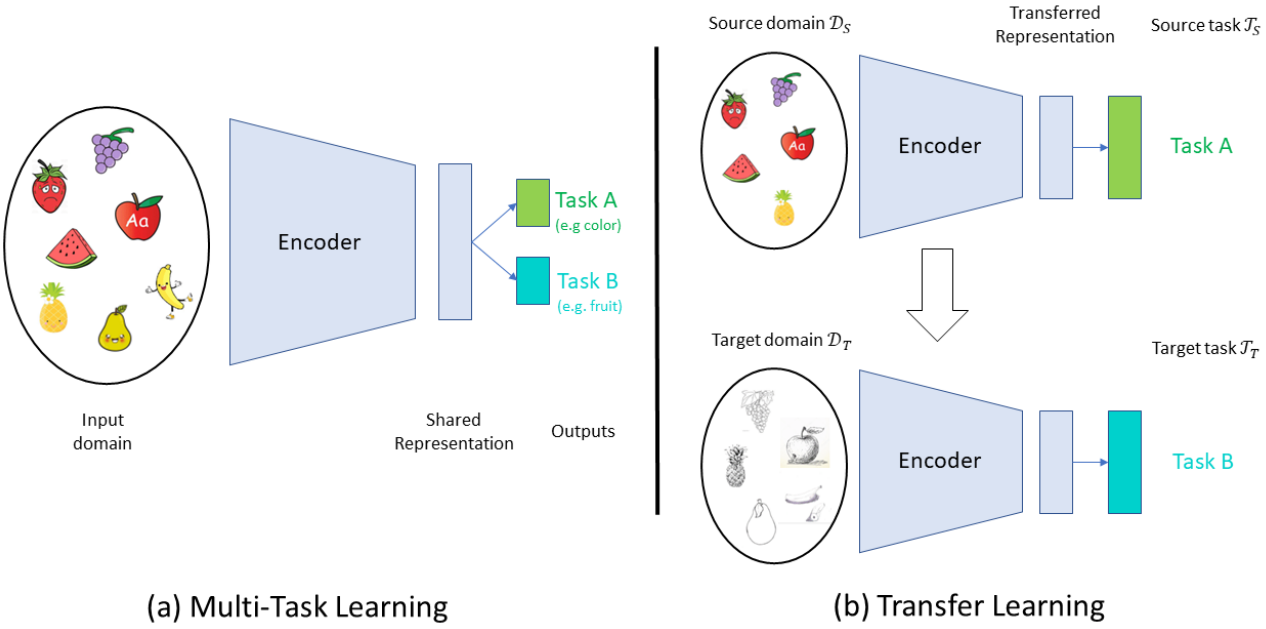


Figure 1.8: (a) Multitask Learning (MTL) consists in learning from multiple tasks simultaneously, assuming that features learned for one task can be re-used for others vs (b) Transfer Learning (TL) in which a model learns from one source domain on a source task and the learned representation is transferred on target domain/task. It also assumes that features learned during pre-training will be re-used during fine-tuning on the target task. Contrary to MTL, source and target domains can be distinct in TL.

hypothesis behind TL is evoked by Y. Bengio et al. [28]: "*intermediate levels of representation [...] can be exploited to share statistical strength across different but related types of examples, such as examples coming from other tasks than the task of interest (the multi-task setting [44]), or examples coming from an overlapping but different distribution*". This hypothesis has been tested later on by J. Yosinski [310] with pre-trained models on ImageNet. They notably found that low-to-middle level features learned on one task can be transferred to others and improve generalization performance. It supports the previous hypothesis that first layers representations are relatively common between tasks (*i.e.,* they share statistical properties) while high-level representations become task-specific.

A more recent work by Neyshabur, Sedghi and Zhang [207] studied this question across different domains (medical imaging with chest X-ray, but also sketches, clipart or painting samples). They showed that **features re-use** plays an important role for transfer learning between tasks *and* domain as well as low-level statistics. In particular, it means that TL boosts generalization performance for images with close visual features than images in source domain. Interestingly, they also showed that pre-trained models make similar mistakes on target domain and have similar representations after fine-tuning on the target task. It means that solutions found after optimization with gradient-descent remain in the same basin of the loss landscape, when using pre-trained models. On the contrary, when trained on the target task from random initialization, final solutions live in different basins and make different mistakes.

Transfer learning in medical imaging. An in-depth analysis has been presented by Raghu et al. [230] on the benefit of transfer learning for medical image classification. They focused their analysis on Chest X-ray and retinal fundus images with large-scale datasets (>200k training examples for each one). Surprisingly, they found no boost in performance when using ImageNet pre-trained models and also no advantage in using large, over-parametrized CNN (compared to lightweights models). On the other hand, they demonstrated convergence speed improvement using pre-training. A finer analysis revealed that features re-use were mostly limited to the first two layers (extracting mostly low-level statistics with Gabor filters). These results suggest that better ImageNet models do not necessarily transfer better for medical images, considering the large discrepancy between source and target domain. More recently, Azizi et al. [15] demonstrated a different trend for big self-supervised models on medical datasets: authors argued that using i) bigger models (*i.e.,* deeper with more parameters) and ii) ImageNet pre-training followed by unsupervised self-supervised pre-training on the target domain both lead to small, but significant, improvement in generalization performance. These two studies indicate the lack of consensus in the scientific community w.r.t TL on medical images. It could be partially explained by the absence of very large-scale medical dataset (like ImageNet for natural images). This issue is particularly present for neuroimaging data and we will come back to it in Chapter 4.

Transfer learning in neuroimaging. Only few works have studied TL on brain imaging data for single-subject prediction (not including segmentation). A recent survey on this topic [283] showed that most of the works tackled classification or segmentation tasks with mostly anatomical MRI data and CNN models. By far, the most studied brain disorder is Alzheimer's disease or, more broadly, neurocognitive impairment. A very complete benchmark on Alzheimer's Disease [304] (AD) showed a small improvement when using Auto-Encoder pre-training for AD detection with 3D anatomical brain images compared to random initialization and poorer generalization with ImageNet pre-training (and more generally any 2D approaches compared to 3D models). This benchmark also pointed out a serious issue: the majority of ML papers reporting results on AD include data leakage during training/test that prevents the scientific community from converging towards a consensus (in particular on the utility of TL in the context of AD detection). Another study on psychiatric disorders [26] demonstrated that brain age prediction pre-training can help to outperform ImageNet pre-training on both AD, schizophrenia, depression, and Mild Cognitive Impairement (MCI) detection. Nevertheless, this study is limited to 2D models (offering generally poorer performance than its 3D counterpart as it does not take into account the 3D spatial structure of the brain) and it does not provide baselines with training from random initialization. The authors argued that their model did not converge in that case.

As we saw through these previous works, deep representation learning allows to ask new

questions about neuroimaging data that could not have be answered with traditional ML, such as: can we learn non-specific features from the healthy population that will reveal new axis of variability for a targeted brain disorder ? Can we learn non-linear relationships from brain anatomical regions to better discriminate brain disorders ? Learning new data embedding with "good" properties (*e.g.,* generalization to unseen data distributions, linear separability between semantic classes, small dimensionnality) is a long-standing goal for deep models. In this thesis we will first start to analyze such models in a supervised context, making the comparison with previous traditional ML models easier. Then, we will focus on techniques specific to deep learning models, in particular unsupervised or weakly-supervised representation learning (only on the healthy population) and transfer learning (from the healthy population to pathological brains on small clinical cohorts).

Chapter 2

Potential and limits of supervised representation learning for neuroimaging

Contents

This work has been submitted to:

Deep Learning Improvement over Standard Machine Learning in Anatomical Neuroimaging comes from Transfer Learning
B. Dufumier, P. Gori, J. Victor, R. Louiset, J-F Mangin, A. Grigis, E. Duchesnay
submitted to NeuroImage 2023

2.1 Introduction

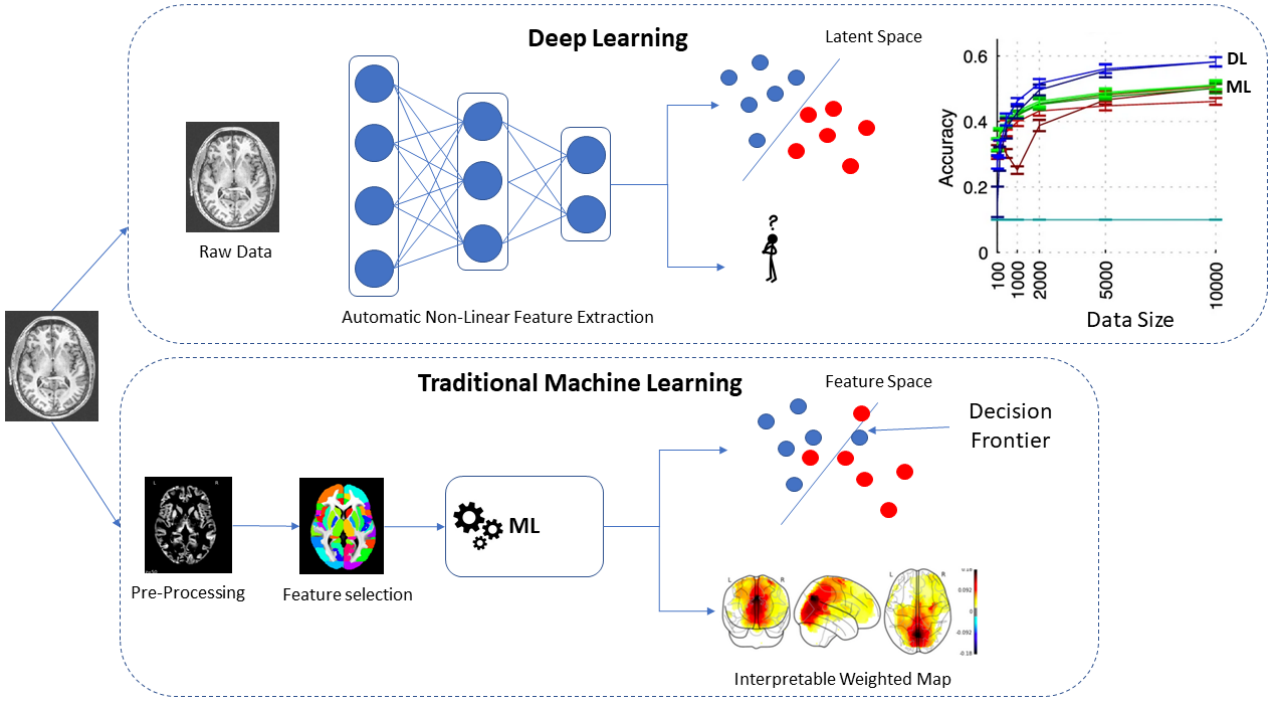


Figure 2.1: Deep Learning (DL) vs "Standard" Machine Learning (SML, that is: linear regression and kernel-SVM) for neuroimaging. DL generally requires no or very little pre-processing and its performance scales very well with increasing sample size for fine-grained classification [190] on ImageNet compared to SML. Do these basic observations on natural images stand for individual-level prediction of mental illnesses and phenotype prediction from brain imaging data?

With the ever-growing availability of brain imaging data (*e.g.*, UK Bioank, HCP, ABIDE, etc.), Machine Learning (ML) and, in particular, Deep Learning (DL) models are starting to emerge for personalized medicine and biomarker discovery in psychiatry and neurology. Psychiatric disorders are complex and highly heterogeneous, gathering both clinical, biological, and environmental variabilities [305], and thus making their neurobiological characterization challenging. In this context, "Standard" ML (SML) models, including (regularized) linear models (such as simple Ridge regression) and kernel-based methods (i.e. Kernel Support Vector Machines [70]), have been broadly used in neuroimaging studies [9, 163], where the number of available samples n is usually small ($n < 10^3$) and the number of imaging features p quite large

(typically $p > 10^5$ for anatomical MRI).

One main drawback that limited their applicability in many medical imaging applications [190] (and more broadly in biomedicine) is their need for pre-selected features manually or automatically designed (*e.g.*, through feature engineering). Specifically, for neuroimaging, the registration and denoising method used, the tissue selected (e.g gray matter and white matter for anatomical data) or the atlas chosen (defining regions-of-interests, a.k.a ROI) for performing the analysis all imply a strong *a priori* that may lower the performance of subsequent ML algorithms used. Moreover, non-linear interactions between input voxels in brain images are not modelled through linear regression and kernel methods provide a simple, yet limited, solution as it is notably sensitive to the "curse of dimensionality" (with poor generalization performance when $n \ll p$).

As opposed to SML methods, DL, and in particular, ConvNets (CNN), can automatically learn from raw imaging data a hierarchical representation of features relevant for the task at hand (*e.g.*, classification or regression). They have shown impressive results on supervised and unsupervised learning problems, both on natural and medical images, by learning a high abstraction of the data in a layer-wise manner. However, as noted in several recent studies [2, 139, 224, 249, 291], the benefit of using DL on anatomical brain MRI data for prediction at an individual level (required for psychiatric disorder diagnosis or prognosis) is unclear, and a careful and extensive comparison with simple regularized linear models and kernel-methods is still missing.

In particular, one worrying observation was made recently [290] for early detection of Alzheimer's disease: as the number of subjects in a study grows, the classification accuracy reported *decreases*. A benchmark on this topic [304] notably confirmed an important bias in the literature due to data leakage during training of ML models, leading to over-optimistic results with small sample size datasets. This may be the case for other tasks in neuroimaging (e.g. schizophrenia detection [246]) and it further justifies the need for a proper comparison of DL models i) at large-scale and ii) with clean cross-validation strategy and independent test sets.

As a result, to answer questions about DL models, we first have pooled a large number of datasets ($n = 19$) across various populations (healthy but also with various brain disorders) and acquisition sites (spread over Europe, Asia and North America). We present this large-scale dataset in section 2.2 before evaluating the representation capacity of supervised DL models in section 2.3 and studying various regularization techniques section 2.4. We consider 3 diagnosis prediction tasks (SCZ vs HC, BD vs HC and ASD vs HC, ordered by task difficulty as measured by ML accuracy) and 2 phenotype prediction tasks (age and sex). We conclude this chapter by showing how uncertainty estimation in DL models matters for both i) increasing their reliability and ii) improving their performance.

2.2 BHB-10K: a large-scale multi-site dataset for transdiagnostic psychiatry

We have gathered a large collection of anatomical brain images to answer key questions with DL models for transdiagnostic psychiatry, including patients with Bipolar Disorder (BD), schizophrenia (SCZ) and Autism Spectrum Disorder (ASD). We build this dataset as a large multi-site database representative of current imaging cohorts available to the research community. We first describe its main statistical properties and then we define our cross-validation strategy to avoid bias associated to the acquisition site where images were acquired (each of which having their own manufacturer and scanning protocol).

2.2.1 Data collection

All data have been collected through various data sharing initiatives, consortium and platforms that can be consulted in the dedicated papers and webpages accessible through hyperlinks shown in Table 2.1. We have reported most of the important demographic information in Table 2.1 for all datasets. Importantly, since we acknowledged that reproducibility is critical for all ML/DL studies, we have also released part the pre-processed data used in this study as a freely available dataset, called the OpenBHB dataset, that can be found here (see Chapiter 4 for a detailed description of this dataset).

The testing splits used for both age and sex prediction are defined using only data from OpenBHB, for reproducibility purposes, as described in section 2.2.2.

2.2.2 Cross-Validation procedure and training splits

For age regression and sex prediction, we have built a multi-site datasets including both OpenBHB (see Table 2.1) - a public dataset that can be accessed without further authorizations- along with more restricted datasets: HCP[286], OASIS 3[184] (only Healty Controls, HC), ICBM[198], BIOBD[241] (only HC), SCHIZCONNECT-VIP[1] (only HC), PRAGUE and BSNIP[271] (only HC). Eventually, we gathered $N = 11210$ scans from 8679 participants and $n = 99$ sites. We first derived an external test dataset with MPI-Leipzig and NAR ($N_{test}^{inter} = 640$ from 619 participants distributed across lifespan from $n = 3$ sites). Then, from OpenBHB, we derived an age/sex/site-stratified internal test dataset and a stratified validation dataset with respectively $N_{test}^{intra} = 662$ scans from 480 participants and $N_{val} = 655$ scans from 482 participants. The remaining training set includes $N_{train} = 9253$ scans from 7098 participants. Importantly, each participant appears in only one split, so that we avoid any data leakage from validation/test set. We chose to use validation/test set only from OpenBHB in order to promote reproducibility in our work[2]. Finally, we sub-sampled this training set in a stratified manner (on age, sex and site) in order to compute performance at varying training sample size

[1] schizconnect.org
[2] OpenBHB is freely available here

	Datasets	Disease	# Subjects	# Scans	Age	Sex (%F)	# Sites	Accessibility
OpenBHB	IXI	HC	559	559	48 ± 16	55	3	Open
	CoRR	HC	1366	2873	26 ± 16	50	19	Open
	NPC	HC	65	65	26 ± 4	55	1	Open
	NAR	HC	303	323	22 ± 5	58	1	Open
	RBP	HC	40	40	22 ± 5	52	1	Open
	GSP	HC	1570	1639	21 ± 3	58	5	Open
	ABIDE I	ASD	567	567	17 ± 8	12	20	Open
		HC	566	566	17 ± 8	17	20	Open
	ABIDE II	ASD	481	481	14 ± 8	15	19	Open
		HC	542	555	15 ± 9	30	19	Open
	Localizer	HC	82	82	25 ± 7	56	2	Open
	MPI-Leipzig	HC	316	317	37 ± 19	40	2	Open
	HCP	HC	1113	1113	29 ± 4	45	1	Restricted
	OASIS 3	Only HC	578	1166	68 ± 9	62	4	Restricted
	ICBM	-	606	939	30 ± 12	45	3	Restricted
	BIOBD [241]	BD	306	306	40 ± 12	55	8	Private
		HC	356	356	40 ± 13	55	8	Private
	SCHIZCONNECT-VIP	SCZ	275	275	34 ± 12	28	4	Open
		HC	329	329	32 ± 13	47	4	Open
	PRAGUE	HC	90	90	26 ± 7	55	1	Private
	BSNIP	HC	198	198	32 ± 12	58	5	Private
		SCZ	190	190	34 ± 12	30	5	Private
		BD	116	116	37 ± 12	66	5	Private
	CANDI	HC	25	25	10 ± 3	41	1	Open
		SCZ	20	20	13 ± 3	45	1	Open
	CNP	HC	123	123	31 ± 9	47	1	Open
		SCZ	50	50	36 ± 9	24	1	Open
		BD	49	49	35 ± 9	43	1	Open
	Total		10882	**13412**	32 ± 19	50	101	

Table 2.1: Demographic information about the datasets used throughout this study. We have gathered 10 openly available datasets to create OpenBHB, from which we have drawn our training set until $N_{train} = 5000$ and our internal and external testing sets for all our experiments on age and sex prediction. We aim at promoting reproducibility of our work by releasing this dataset pre-processed to the neuroimaging community. You can find a first version here.

($N \in [100, 500, 1000, 3000, 5000, 9253]$) for both age and sex prediction using a Monte-Carlo Cross Validation (CV) procedure, similarly to [2, 249]. We repeated this sub-sampling 5 times for $N \leq 500$ and 3 times otherwise in order to keep a reasonable computational budget, while still deriving a consistent estimator of classifiers performance. About schizophrenia, bipolar disorder and autism detection, we detailed the splits used in Table 2.2. We used the same splits for all models (SML and DL) and we repeated each experiment 3 times, using different random initialization, reporting the average and standard deviation.

2.2.3 VBM and Quasi-Raw pre-processing

VBM pre-processing is performed with CAT12 [109] from the SPM toolbox. It essentially consists in noise and bias-field correction followed by Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF) segmentation. Images are non-linearly aligned to the MNI template with DARTEL[12] and modulated using the Jacobian map of the deformable transformations.

Task	Split	Datasets	# Subjects	#Scans	Age	Sex(%F)
SCZ vs HC	Training	SCHIZCONNECT-VIP, CNP PRAGUE, BSNIP, CANDI	933	933	33 ± 12	43
	Validation		116	116	32 ± 11	37
	External Test		133	133	32 ± 12	45
	Internal Test		118	118	33 ± 13	34
BD vs HC	Training	BIOBD, BSNIP CNP, CANDI	832	832	38 ± 13	56
	Validation		103	103	37 ± 12	51
	External Test		131	131	37 ± 12	52
	Internal Test		107	107	37 ± 13	56
ASD vs HC	Training	ABIDE 1+2	1488	1526	16 ± 8	17
	Validation		188	188	17 ± 10	17
	External Test		207	207	12 ± 3	30
	Internal Test		184	186	17 ± 9	18

Table 2.2: Training/Validation/Test splits used for the 3 mental illness disorders detection. The external test set is always made by out-of-site images and each participant falls into only one split, avoiding data leakage. The internal testing set is always stratified according to age, sex, site and diagnosis, as well as the training and validation set. All models use the same splits.

All sMRI scans are re-sampled to have an isotropic 1.5mm^3 spatial resolution with dimension $121 \times 145 \times 121$ using a linear spline interpolation. Going to higher spatial resolution would have induced a bigger computational burden and considering the difference in scanner parameters in our cohorts (e.g., permanent magnetic field), we decided to fix this resolution for all images. We also normalized all images using the Total Intracranial Volume (TIV) estimated by CAT12 to account for the (irrelevant) differences in head size.

As opposed to VBM, quasi-raw pre-processing was designed to be minimal. Only essential steps have been kept in order to map the images coming from different sites and scanners to the same space with the same resolution and only important image correction steps have been applied. Specifically, each scan is rigidly re-oriented to the MNI space and then re-sampled to a 1.5mm^3 spatial resolution through a linear spline interpolation. The bias field is corrected using the N4ITK algorithm [282] from ANTs [13] and the brain is extracted with BET2 [159] (the skull and non-brain tissues are removed). Each image is linearly registered (9 degrees of freedom) to the MNI template with FLIRT from FSL [158].

For all pre-processed images, we applied a visual quality check and we removed images poorly segmented or with obvious MRI artefacts.

2.3 Representation capacity of supervised deep models at scale

Can DL models exploit non-linear relationships from brain images to predict individual phenotypes and mental illnesses ?

In a recent study [249], Schulz et al. studied whether the two main priors encoded in current CNN, namely translational equivariance (derived from the convolution operation) and compositionality (derived from its hierarchical structure), can be exploited to capture non-

linear dependencies in structural/functional Magnetic Resonance Imaging (sMRI/fMRI) data for individual prediction tasks with UK Biobank [37] (UKB). In particular, they showed that linear and DL models have a similar scaling trend, even in the large-scale regime ($N_{train} = 8$k), on both modalities (sMRI and fMRI) for a variety of tasks (age and sex prediction but also fluid intelligence or household income prediction). It notably suggests the incapacity of DL models to learn non-linear functions on brain images. They proposed that current noise in these data prevent DL from outperforming simple linear models. It was notably exemplified on MNIST where CNN matches linear model performance when sufficient Gaussian noise is added.

However, their results directly contradict the ones obtained by Peng et al.[218] on UKB for brain age prediction, as noted by Abrol et al. [2]. Specifically, they pointed out some technical flaws in the work of Schulz et al. that drastically affect their conclusions. The main shortcomings were the feature selection step performed for both SML and DL (with an arbitrary number of reduced dimensions) and the use of a single central brain slice in their main experiments, which limited DL representation capacity. On the contrary, Abrol et al. showed a significantly better scaling trend for DL on UKB with training samples ranging from $N_{train} \geq 2000$ to $N_{train} = 10^4$ when feature selection were only for SML models, and they used a whole-brain approach for DL. They attributed the performance drop between Schulz et al. and Peng et al. to a coding bug. Moreover, they also found a small but significant increase in performance on the Mini-Mental State Examination (MMSE) regression task ($N_{train} = 428$, -0.07MAE, Mean Absolute Error, for DL vs. SML) on the ADNI dataset [156] (comprising a population of Alzheimer patients), which might be in contradiction with a recent benchmark [304] on Alzheimer's detection that found no significant differences between SML and DL. While this score represents an indicator of Alzheimer's disease severity, it does not translate into Alzheimer's diagnosis [81], which may explain the different findings. Finally, they showed that DL is capable of extracting robust interpretable brain representations, even in the small data regime for MMSE regression task, consistently across runs and saliency methods.

However, the studies of Abrol and Schulz provide only a partial analysis about the DL capacities for neuroimaging data that we aim to extend in this work.

First, most recent papers [2, 218, 249] have mainly focused their analysis on phenotype prediction in the healthy population, including socio-demographic and lifestyle measures. While studying phenotype prediction has become an important research field for many research questions (new biomarkers discovery for psychiatric disorders or neurocognitive impairment with brain age [64, 68, 164, 178] or normative modeling [197, 305, 313]), fair DL evaluation on psychiatric disorder classification is (also) urgently required. The question of whether non-linearities can be captured in highly heterogeneous clinical cohorts including patients with schizophrenia [178, 305] (SCZ), bipolar disorder [305] (BD) and autism spectrum disorders [313] (ASD) is still debated, and no clear consensus arises [224, 240, 304], mainly because of the small sample size of the current datasets (typically $N < 10^3$) which causes ML models to over-fit and bias the

neuroimaging community towards over-optimistic results [99, 165, 223, 246]. These pathologies involve subtle anatomical atrophies/hypertrophies in cortical and subcortical structures, and their identification in a case-control manner is still a difficult challenge.

Second, both Abrol et al. and Schulz et al. have based their analysis mostly on a unique homogeneous (*i.e.* single-scanner model) dataset (UKB), that does not reflect the inevitable heterogeneity in emerging large multi-site clinical data collections (*e.g.*, ABIDE, ABCD, SCHIZCONNECT, etc.). As such, a comprehensive complementary benchmark on phenotype prediction with large-scale multi-site datasets is required. As noted by Koppe et al. [176], since DL has an exceptional capacity to learn any function (even random noise [317]), it can also learn "disease-irrelevant site-specific characteristics," and its generalization capacity on data acquired on never-seen sites must also be reported.

2.3.1 Deep learning vs good old Tikhonov regularization

First, in our study we analyze the scaling trend of several DL architectures on age and sex prediction in the healthy population using BHB-10K. Our experimental setting has several key differences with the current literature: i) we apply no feature selection strategy on *both* DL and SML, as we observed a strong degradation in performance with the experimental design previously used in [2, 249]; ii) we separately predict age and sex in order to avoid arbitrary age discretization ; iii) we assess the generalization performance on both an external test set ($N_{test}^{inter} = 640$), including never-seen sites, and an internal test set ($N_{test}^{intra} = 662$) stratified on age, sex and site. The use of an external test site should prevent the model from over-fitting on confounding variables related to the site-specific information [293].

As for brain disorder detection, we train each DL classifier with a binary cross-entropy loss, treating each task as binary classification. Importantly, these three tasks do not have the same difficulty (at least w.r.t their accuracy score [92, 240]), and one might expect improvement with non-linear models on harder tasks where SML models under-perform (e.g., autism).

We chose three DL models representative of the current SOTA for both computer vision and neuroimaging tasks [1, 2, 87], namely AlexNet (corresponding to DL1 in study [181], 2.5M parameters, the smallest with only 5 convolutional layers), ResNet18[135] (33.2M parameters) and DenseNet121 [149] (11.2M parameters, the deepest model among the three chosen with 121 layers). Importantly, we adopted a 3D architecture for each of these networks in order to account for the 3D spatial structure of our images. It means that we adopted 3D filters in each convolutional layer. AlexNet performed on par with current SOTA on age prediction (SFCN [218]) and it allows us to be comparable with the recent literature on phenotype prediction and MMSE regression task. Increasing the depth of DNN also provides interesting insight into the complexity of the models required at large-scale on brain imaging data.

We compared their performance and generalization power against two regularized linear models (only ℓ_2, *i.e.,* ℓ_2-regularized/logistic regression for regression/classification, or $\ell_1 + \ell_2$

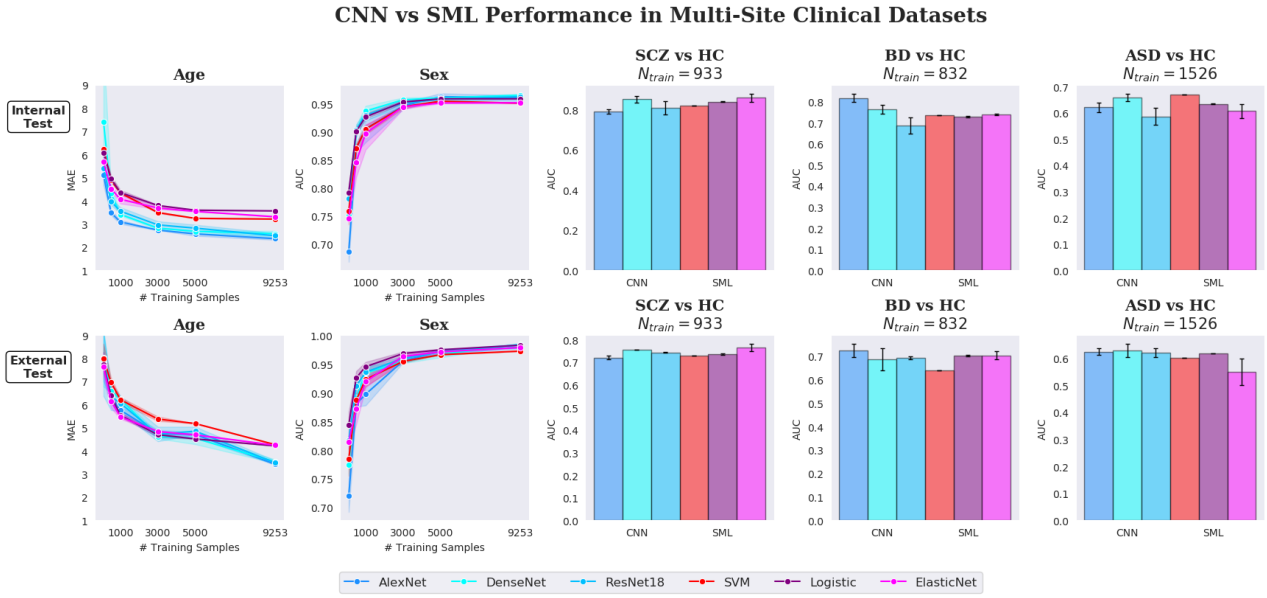


Figure 2.2: DL vs. SML performance on phenotype prediction and increasingly difficult diagnosis classification tasks on highly multi-site datasets. For SML methods, 2 linear models with ℓ_1 (Logistic) or $\ell_1 + \ell_2$ (ElasticNet) penalization are evaluated as well as non-linear Radial Basis Function (rbf) SVM. As for DL, vanilla AlexNet[181] (previously introduced by Abrol et al.[2]) and more advanced ResNet18[135] and DenseNet121[149] (121 layers taking advantage from skip-connections and feature re-using) are considered. Importantly, both DL and SML algorithms are trained on whole-brain 3D anatomical images. All models are evaluated on two different test sets: an internal test stratified on age, sex, site ($N^{pheno}_{test} = 662$, $N^{pheno}_{val} = 655$), and diagnosis for clinical cohorts ($N^{scz}_{test} = 118$, $N^{scz}_{val} = 116$, $N^{bd}_{test} = 107$, $N^{bd}_{val} = 103$, $N^{asd}_{test} = 184$, $N^{asd}_{val} = 188$); an external test including sites never seen during training ($N^{pheno}_{test} = 640$, $N^{scz}_{test} = 133$, $N^{bd}_{test} = 131$, $N^{asd}_{test} = 207$). Models cannot use any site-specific information for their prediction on this test set, eliminating a strong bias reported in the literature. For age and sex prediction, we performed 5-fold (resp. 3-fold) Monte Carlo Cross-Validation sub-sampling procedure for $N_{train} \in \{100, 500\}$ (resp. $N_{train} \in \{1000, 3000, 5000, 9253\}$). As for diagnosis classification tasks, each model is trained 3 times with different random initialization and average and standard deviations are reported. Mean Absolute Error (MAE) is the reference measure for age prediction while Area Under the Curve (AUC) is the preferred metric for binary classification tasks since it does not depend on a particular threshold (it only measures a classifier discriminative power). Overall, SML models perform equally well with DL models for sex prediction (up to $N_{train} = 9253$), SCZ vs HC, BD vs HC and ASD vs HC. Both SML and DL performance keeps improving for age prediction when increasing the number of training subjects N_{train} on the external test. On the other hand, performance increases very slowly (it is almost a plateau) on the internal test starting from $N_{train} \approx 3k$ with an important improvement for non-linear DL models over SML.

i.e., ElasticNet) and one non-linear (Radial Basis Function kernel) rbf-SVM, that showed good performance for both psychiatric disorders and neurodegenerative disease [240, 251]). Results are based on a Monte-Carlo Cross Validation strategy as detailed in section 2.2.2. In order to fairly compare both DL and Standard Machine Learning (SML, including linear models and Kernel-SVM), we perform these experiments on VBM data. Indeed, all images are non-linearly registered to the same template so that each voxel contains information from the same spatial location between different subjects.

From Fig. 2.2, we observe very similar performance on all classification tasks (both sex prediction and diagnosis classification) across all models and even in the very large data regime ($N_{train} > 9000$ for sex prediction). Specifically, all models achieve almost perfect AUC score

(Area Under the Curve) on sex prediction on both test sets (AUC = 98.32 for Logistic Regression and AUC = 98.47 for DenseNet with $N_{train} = 9253$ on the external test set). While DenseNet is almost always the best performing network for detecting schizophrenia, bipolar disorder, and autism, it achieves performance on par with Logistic ℓ_2 and rbf-SVM, i.e $\approx 85\%$ AUC on SCZ vs. HC, $\approx 76\%$ AUC on BD vs. HC and $\approx 65\%$AUC on ASD vs. HC, on the internal testing set, losing resp. -10%AUC, -8%AUC, and -3%AUC on the external test set. A similar trend can also be observed for the other models. This suggests that DL fails to capture additional information with respect to linear model, such as highly non-linear dependencies, possibly due to large noise in the input data [249] and high inter-individual heterogeneity in neuroanatomical images [210, 305, 313].

Interestingly, we observe a different trend for age prediction. DL models are more accurate than SML on both test sets, with a significant improvement even from $N_{train} \geq 1000$ on the internal test set (ΔMAE = 0.98, $p < 0.0012$ between AlexNet and ElasticNet with $N_{train} = 1000$). DL performance on the external test set is also significantly better than SML but it needs much more training samples (ΔMAE = 0.82, $p < 10^{-5}$ between AlexNet and ElasticNet with $N_{train} = 9253$). This gain in performance has been reported in several recent studies [2, 218] and it contrasts with the results on psychiatric disorders.

We also remark that we reach SOTA performance on age prediction as compared to previous studies [2, 218] on this topic (with MAE=$2.36_{\pm 0.04}$ on the internal test[3]), which also validate the choice of the architecture designs for DL models.

The discrepancy of results between internal and external test (with a constant and significant decrease in performance for all models) is interesting to notice. It notably suggests a high overfitting issue for both DL/SML on acquisition site. This recurring issue has been reported in the literature (e.g. Alzheimer's detection [304] or demographic factor prediction [293]) and may explain the high variability of performance reported in the literature on these tasks.

Our evidence on psychiatric disorder classification (but also sex prediction) support the main hypothesis made by Schulz [249, 250]: "high levels of noise in neuroimaging data may effectively linearize decision boundaries, potentially leaving little nonlinear structure for machine learning models to exploit". Furthermore, as noted by [209] on functional MRI (but transposable to structural imaging), spatial averaging over $\approx 10^4$ neurons in each voxel and small sample size may also play they part as it can easily linearize macroscopic brain dynamics.

2.3.2 Do deep models benefit from raw data ?

In the previous section, we show how scaling trend of DNN were similar to that of linear models on anatomical imaging. However, we emphasize that we used highly pre-processed VBM images including only gray-matter volume measure in each voxel as input data. These images were non-linearly registered to a template, meaning that the actual folding patterns were largely

[3]We emphasize that, even if the data size is comparable with previous works, it is not a direct comparison since previous studies used a different test set stratified on UKBioBank.

removed.

As pointed out by Y. Lecun, Y. Bengio and G. Hinton [190], DNN excels at learning from raw images, by performing automatic feature extraction for pattern recognition. On the other hand, recent findings on brain age prediction [67, 152, 218, 293] suggest that DL models perform similarly between raw images (with only linear registration and eventually non-brain tissue removal) and fully pre-processed ones (with non-linear diffeomorphic registration, gray matter extraction, and several bias correction steps), suggesting that DNN do not extract extra-information from raw data. This is a major difference with classical vision tasks (e.g., ImageNet classification) since we know that automatic feature extraction of color, shape, and texture is the cornerstone of today's CNN performance. As a result, a fundamental question is whether usual non-linear computationally demanding pre-preprocessing steps actually remove non-linear discriminative information for brain disorders that could have been leveraged by DL (e.g., cortical folding patterns). This problem has not been addressed for mental disorders such as schizophrenia, bipolar disorder, and autism.

Scaling trend and over-fitting effect

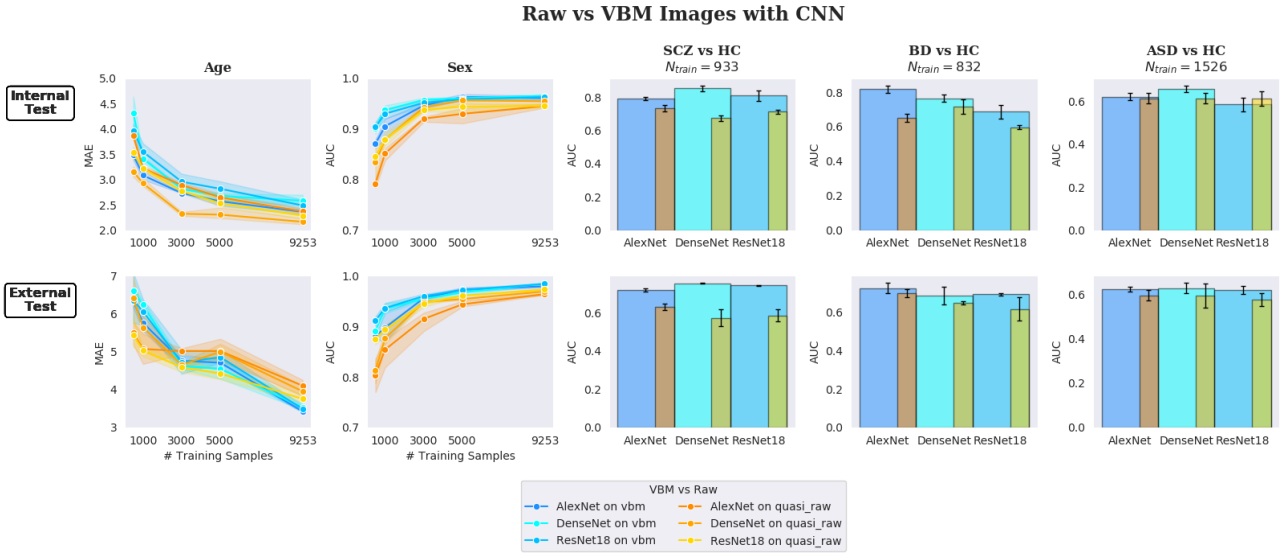


Figure 2.3: DL performance are evaluated on both (quasi) raw brain images and extensively pre-processed, non-linearly registered, anatomical Gray Matter (GM) brain images (namely VBM). As before, three CNN families (AlexNet, ResNet18, DenseNet121) are trained on increasingly large training sets for age and sex prediction and three diagnosis classification tasks. They are tested on both the internal (stratified) test set and the external one (including sites never seen during training). DL models fail at extracting more discriminative features from raw brain images than fully pre-proprocessed ones, even in the large-scale data regime. This observation contrasts with their exceptional automatic feature extraction capacity on natural images.

We take the same setting as previously but we replace VBM images by their quasi-raw counterpart (i.e. with minimal pre-processing). In particular, it means we preserve both gray matter, white matter and CSF signal as well as the geometry of folding patterns (e.g. curvature, depth, etc.) . We perform only limited noise reduction and we refer the reader to section 2.2.3

for more details. As before, we report learning curves for phenotype prediction and we use only the maximum number of available samples for psychiatric disorders.

Surprisingly, from Fig. 2.3, we observe that, globally, CNN do not perform any better on raw T1 scan than on VBM data, at least on the external test. More specifically, we observe a degradation of performance of 1.6%AUC for sex classification and of 0.25 MAE, $p < 0.05$, for age regression with $N_{train} = 9253$ with DenseNet and ResNet respectively, the best performing models on these two tasks on the external test set. About the classification of psychiatric disorders, this effect is even more pronounced with -14%, -4% and -3% AUC on average between performance on VBM and raw data for schizophrenia, bipolar disorder, and autism respectively on the external test set. Interestingly, while these observations are confirmed on both internal and external test set for all psychiatric disorders and sex prediction, we do not observe the same trend for age prediction (again) between the internal and external test set: CNN seem to over-fit more on sites, showing much worse performance when testing on a site-independent cohort.

To explain these intriguing results, we hypothesize that raw measurements induce much more noise in the signal (especially related to acquisition site), leading to even poorer results than VBM (even if the raw data contains theoretically much more discriminative signal). Again, this favors the hypothesize by Schulz from another perspective. We intend to check this hypothesis in the next section.

Towards a first explanation: raw images overwhelmed by site-related noise

Pre-processing		SCZ vs HC	BD vs HC	ASD vs HC
VBM	Site Pred.(%)	$29.07_{\pm 3.73}$	$26.43_{\pm 2.07}$	$7.01_{\pm 1.53}$
Raw	Site Pred.(%)	$70.71_{\pm 3.36}$ (+41%)	$82.92_{\pm 3.86}$ (+56%)	$48.74_{\pm 5.88}$ (+41%)
	Random Level	10.0	7.69	3.45
	Δ AUC=VBM-Raw	14%	4%	3%

Table 2.3: Site prediction balanced accuracy (in %) from latent representation of DenseNet trained on psychiatric disorder classification. We reported the random level when predicting random sites ($= 1/n_{sites}$) as well as the difference ΔAUC between performance on psychiatric classification from VBM and raw data. It clearly shows a much higher over-fitting effect on site (viewed as noise) for raw data compared to VBM even when the model is not trained on this task. This could be a partial explanation for the drop in performance between VBM and raw data.

To check this hypothesis, we first plotted both quasi-raw and VBM pre-processed images (from internal and external test set) encoded by a DenseNet trained on age prediction with $N_{train} = 9253$ (see Fig. 2.4). We used t-SNE [284] visualization to map the embedded images to 2D representations. We observe a clear difference, in the embedded space, between raw images coming from either the internal or external test set (especially for middle-aged participants between 20 and 40 years old). This is clearly not the case for VBM images, where both inter- and intra-site images correctly overlap in the embedded space for a given age range (blue/range

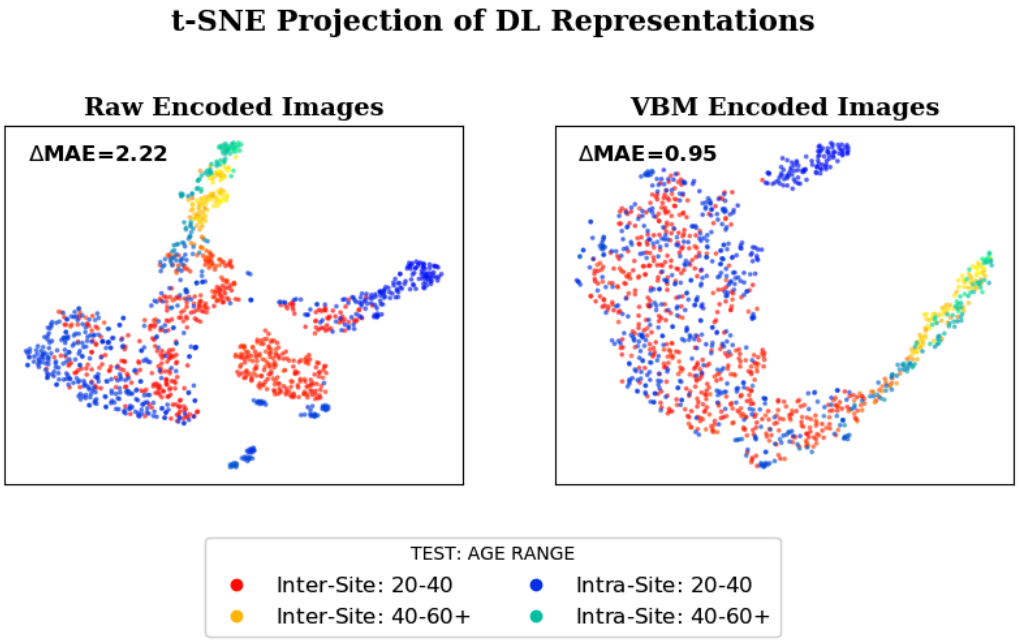


Figure 2.4: t-SNE visualization of raw vs VBM images encoded by DenseNet trained on age prediction with $N_{train} = 9253$. We distinguished images from internal test (coming from already-seen sites) and external test. Here ΔMAE = |MAE(external test) − MAE(internal test)| where MAE(x) corresponds to the age prediction MAE (Mean Absolute Error) for the test set x. It can thus be seen as a proxy to measure the domain gap between internal and external test sets. Distinct regions for the same age range (blue/red and yellow/cyan) can be observed when encoding raw images. However, these regions clearly overlap for VBM encoded images. It suggests a higher over-fitting effect related to site on raw images than on VBM.

and yellow/cyan). This greater difference (i.e., domain gap) between internal and external test sets for raw encoded images could explain the differences shown in Fig. 2.3 for age prediction, supporting the site over-fitting hypothesis.

Furthermore, we make an indirect test to check whether noise induced by the scanner explains the discrepancy in results between VBM and raw measurements on psychiatric disorders. From DenseNet trained to predict a given psychiatric condition with a given pre-processing (VBM or raw), we train a linear classifier to predict acquisition site from the network representation. Specifically, we train a linear classifier to predict acquisition site on top of the penultimate layer of DenseNet trained to predict psychiatric condition. Importantly, DenseNet's weights are frozen so its representation is fixed. We have reported the balanced accuracy obtained on site prediction task in Table 2.3. In Table 2.3, we notably show an increase $> 40\%$ in balanced accuracy (Bacc) on site prediction when the network is trained on raw data rather than VBM to classify psychiatric conditions. From an information bottleneck point-of-view, it suggests that the network fails at compressing disease-related features from raw images and rather tends to rapidly over-fit on scanner-induced noise .

In conclusion, these evidence support our hypothesis that raw measurements contain too much noise that prevent DNN from learning non-linear boundaries and, overall, it degrades the downstream performance even compared to fully pre-processed images. Even if evidence

show that folding patterns are predictive of psychiatric disorders (e.g. increased gyrification index during childhood for ASD and during adolescence for schizophrenia, see [242] for a recent review), DNN seems to fail at exploiting such complementary information buried inside raw measures. As suggested by Schulz [250], more anatomical prior information needs to be integrated during learning. We will dig into that lead in the next chapiter.

2.3.3 A closer look at deep models with brain region importance analysis

While DL models are often considered as a "black box", several interpretability methods have been proposed over the years to highlight the discriminative image areas used by the model to take its decision (see this recent survey by Zhang et al. [319]). Here, we aim at discovering whether DL and linear models take their decision based on the same brain region patterns, which is a critical question for precision psychiatry. If two models strongly disagree on the discriminative power of the same brain area, which one can we trust ?

In this regard, linear models are much simpler to interpret since we have direct access to the weighted maps (also called "importance maps" [20]). In a weighted map, each weight is associated to a unique input feature. Higher absolute weight values indicate a stronger importance of the corresponding input features on the final prediction score. In particular, in a clinical context with anatomical images, hypertrophy (resp. atrophy) in regions with high positive (resp. negative) weights translates into a stronger brain signature for a given pathology, i.e a higher predictive score.

As a generalization to the non-linear case, we have chosen a gradient-based method [259] for DL model interpretability. This sensitivity analysis computes the gradient of predicted output w.r.t. each input voxel (*i.e.*, it quantifies how much output prediction varies with each input voxel). More sophisticated gradient-based models have been proposed over the years, but they do not necessarily result in more accurate saliency maps [3]. Similarly to Abrol et al. [2], we compute brain region importance maps using the Automated Anatomical Labeling atlas [236] (AALv3) containing 166 parcellations. Specifically, for each input image, a weighted map is computed through sensitivity analysis and all absolute values are summed per region. The resulting importance map is normalized so that it sums to one. Finally, all importance maps for each test set (internal and external) are averaged. We reported these maps on the external test for visualization purposes in Fig. 2.3.3. Importantly, all models used for computing importance maps are trained with the maximum number of training samples (which is the best-case scenario).

To easily compare region importance obtained with linear and DL models, we have computed the correlation matrix between all averaged maps in Fig. 2.5.

Fig. 2.5 shows two clear patterns, both reproducible across testing set. First, all DL models use the same cortical and sub-cortical areas to take their decision. Similar saliency maps are obtained between DL and logistic regression with ℓ_2 regularization for all tasks (correlation

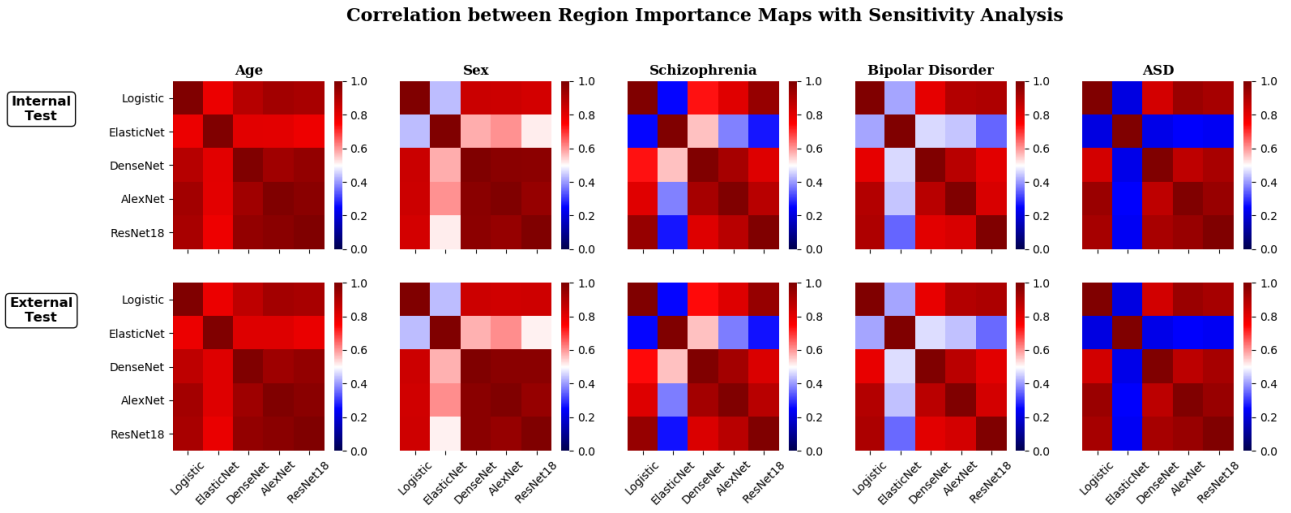


Figure 2.5: Correlation matrix computed between brain region importance maps obtained for each task and model. Strong correlation indicate a good agreement between two models for a given task. Each brain region importance map is obtained through sensitivity analysis (i.e using a gradient-based method) for both DL and linear models. All models considered have been trained with the maximum number of training samples. Brain regions are defined through the AAL atlas, similarly to[2].

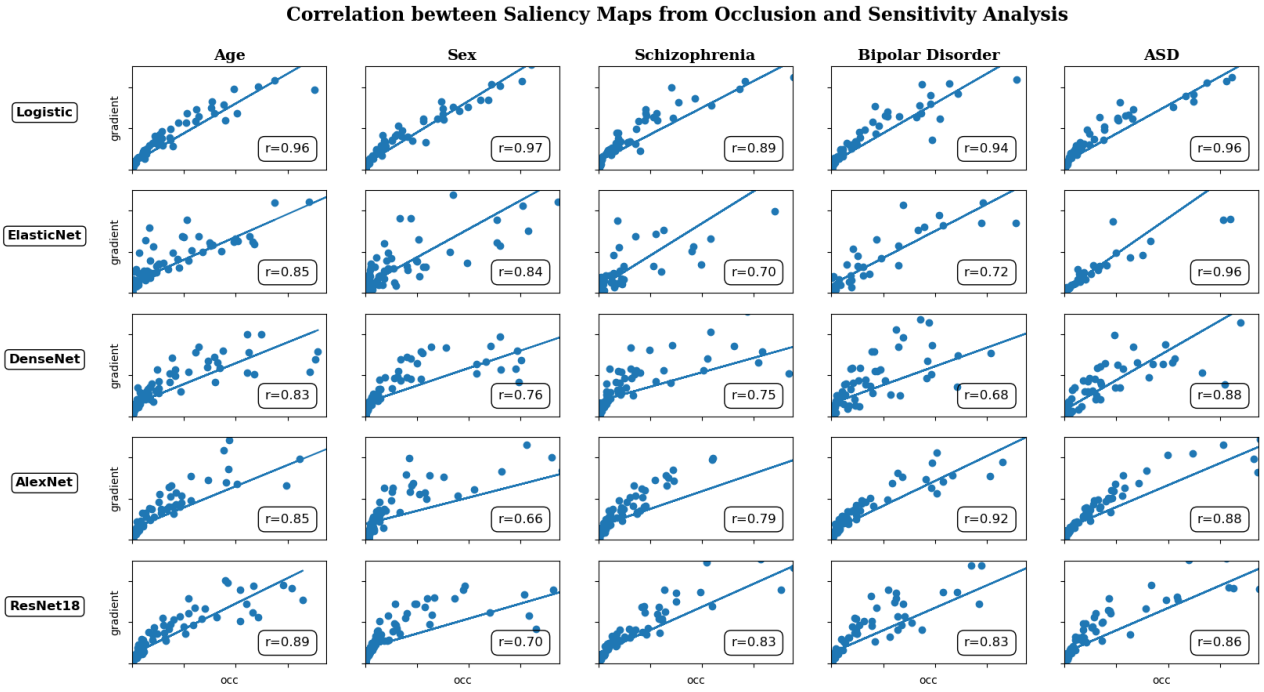


Figure 2.6: The correlation between saliency maps obtained from occlusion and sensitivity analysis are reported for all models and tasks.

$r > 0.70$ between the linear model and all DL models for all tasks). This is in line with recent studies [20, 240] on SML models applied to age prediction, schizophrenia, and bipolar disorder detection. Both linear and non-linear models resulted in similar final weighted maps, with various degrees of noise and sparsity. Second, ElasticNet generates extremely sparse maps (which is expected) but with regions overall poorly correlated with other models ($r = 0.21$, $r =$

0.22, $r = 0.25$ and $r = 0.24$ between ElasticNet and Logistic ℓ_2, DenseNet, ResNet and AlexNet resp. on ASD detection). Overall, this is more pronounced as we increase the task difficulty (e.g., age or sex prediction with $> 95\%$AUC vs. ASD detection with $\approx 60\%$AUC). We may be tempted to relate these poor correlations directly to the relatively small sample size in clinical cohorts than for phenotype prediction ($N_{train} < 2000$ for the former vs $N_{train} \approx 10$k for the latter). Nonetheless, we observe a rather good correlation for schizophrenia detection between DenseNet and ElasticNet (around 60%), while, for ASD, all correlations are low ($r < 30\%$). It suggests a higher inter-individual heterogeneity in cortical discriminative patterns for ASD compared to schizophrenia, which would explain i) poor performance and ii) high variability in saliency maps. Zabihi et al. [313] notably showed how cortical thickness (CT) alterations differ from one sub-group ASD population to another, even for match ages (e.g. decrease CT during childhood vs increase CT for some patients in other areas). Our saliency maps analysis may notably highlight the high biological variability for ASD, reflecting the fuzzy boundary delimiting this pathology based on DSM-5 criteria [203].

Finally, since this experiment only relied on sensitivity analysis, we have validated our methodology using an occlusion-based method [315]. Occlusion essentially consists in monitoring the model prediction variation while occluding each brain region independently (defined by the AAL atlas in our case). As before, we performed this analysis for all models and tasks (since occlusion is model-agnostic) and we have reported in Fig. 2.6) the correlations between the saliency maps obtained from occlusion vs. sensitivity analysis. Overall, we found an excellent agreement between these two methods ($r > 0.70$ for all models and tasks except AlexNet with sex prediction and DenseNet on bipolar detection). This comforts our previous observations although we acknowledge that a finer analysis on saliency maps at the individual-level may reveal much more inter-model differences than our group-level analysis.

2.4 Model regularization and data harmonization

DNN can generalize very well to unseen natural images when trained on a sufficiently large and representative bank of images. This assertion is true at least on standard vision tasks (involving object classification on ImageNet [74] or segmentation for instance) on which humans are also very good at and can easily perform. Generalization means that the gap between training and test error is small even (and especially) when the number of parameters is extremely large compared to the number of training examples. Theoretically, DNN should be able to overfit perfectly all the training set, leading to very poor generalization error. Zheng et al. [318] notably show that current SOTA DNN can very well fit random labels (on exactly the same "standard vision datasets" as aforementioned), demonstrating that most mathematical tools currently used to explain DNN generalization power should be rethought (e.g. Rademacher complexity, VC-dimension etc.)

Nevertheless, in practice, when trained with stochastic gradient descent, DNN prefer to

extract semantic information from images to perform their task (that is, high-level meaningful features that we-as humans-also use). In previous section, we saw that the story was different for neuroimaging data: DNN appear to over-fit very well and rapidly on the training set, only matching the performance of linear models. Over the years, several regularization methods have been invented to limit such over-fit and improve their generalization power. Data Augmentation is certainly one of the most famous and, according to Zheng et al [317] and confirmed in [141], the most efficient regularization technique (compared to classical dropout [265] and weight decay). Inspired by the human perception, it became the crucial component of today's most effective self-supervised and semi-supervised models (e.g. SimCLR [52] and FixMatch [262], see Chapter 3 for a thorough discussion). Here, we evaluate the capacity of data augmentation on neuroimaging data and draw first conclusions and concerns about its utility.

2.4.1 Data augmentation as regularization: myth vs reality

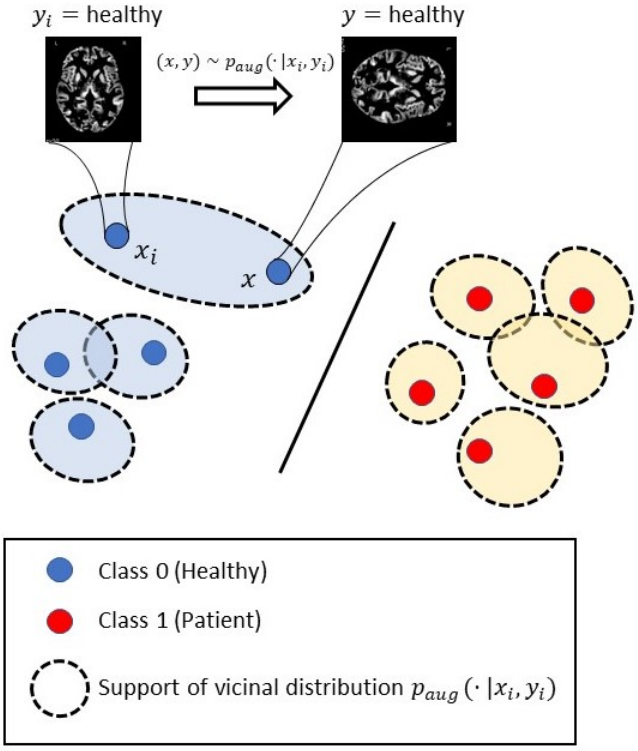


Figure 2.7: Data augmentation as vicinal distribution sampling. From a given labelled image (x_i, y_i), augmented images (x, y) are generated from a vicinal distribution $dP_{x_i}(x)\delta_{y_i}(y)$ corresponding to the augmentation module (e.g. geometrical transformations such as image rotation or cropping, cutout [76], etc.) Here, we assume the images generated has the same label y_i as the original image but this assumption can be relaxed (e.g. Mixup [318]). Deep models are trained on these generated images, learning from a much larger and diverse set of images (covering a broader region in the input space).

When working with rather small-scale data-sets (typically $N \approx 1$k) and large input images (> 1M voxels), data augmentation offers a simple way to artificially increase the dataset size by applying transformations on training images to generate a larger and more diver set of labelled

images (see Fig. 2.7). From the Vinicial Risk Minimization (VRM) point-of-view, Chapelle et al. [48] shows that it can be seen as a regularization technique that imposes invariance to given transformations for a prediction task (we detail it below). More profoundly, it has been suggested that applying data augmentation during training lead to more biologically plausible representations inside DNN [142], as it robustify the network against identity preserving image transformations (a property already observed in the human medial-temporal lobe [225]).

We first describe theoretically data augmentation based on vicinal risk minimization to justify the transformations used. Then, we provide our empirical study on mental illness disorders (where DNN currently fails at extracting non-linear relationships) and phenotype prediction (where we successfully show an improvement of DNN over linear models- at least for age regression).

Vicinal Risk Minimization

In a supervised learning problem, we aim at learning a function $f \in \mathcal{F}$ that maps input data $x \in \mathcal{X}$ (e.g. image) to label $y \in \mathcal{Y}$ (e.g. human annotation). The relationship between x and y is modelled as a joint distribution P from which (x, y) is sampled. In a real-world setting, f is trained on a limited number of examples thanks to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ penalizing the difference between the predicted label $f(x)$ and the true one y. The risk of f is defined as:

$$\mathcal{R}(f) = \mathbb{E}_P \ell(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(x), y) dP(x, y) \tag{2.1}$$

In practice, P is unknown but we have access to n examples $(x_i, y_i)_{i \in [1..n]} \sim P$ to approximate the risk $\mathcal{R}$. A standard approach consists in defining the empirical joint distribution:

$$d\hat{P}_\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x) \delta_{y_i}(y) \tag{2.2}$$

where δ_x is the Dirac mass function centered at x. Plugging this estimate in eq. 2.1 gives the empirical risk estimator:

$$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \tag{2.3}$$

Minimizing this empirical risk for supervised learning is known as Empirical Risk Minimization (ERM) and was formalized by Vapnik [288] in 1999. Nevertheless, the main issue with ERM is the risk of over-fitting and under-fitting, depending on the class of functions $\mathcal{F}$ considered. This notably conditions the generalization guarantee of the model f as we briefly mentioned above. Chapelle et al. [48] proposed to replace δ_{x_i} in eq. 2.2 by another estimator of the distribution in the *vicinity* of x_i, $dP_{x_i}(x)$. It notably induces a new empirical vicinal distribution:

$$d\hat{P}_{vic}(x, y) = \frac{1}{n} \sum_{i=1}^{n} dP_{x_i}(x) \delta_{y_i}(y) \tag{2.4}$$

From $d\hat{P}_{vic}(x, y)$, it is possible to define the empirical vicinal risk as:

$$\hat{\mathcal{R}}_{vic}(f) = \frac{1}{n} \int \ell(f(x_i), y_i) dP_{x_i}(x) \tag{2.5}$$

The advantage of Vicinal Risk Minimization (VRM) over ERM becomes clear with this formulation: if the class of functions $\mathcal{F}$ is not well-suited for the task (i.e. too much capacity conducing to rapid over-fit) then a better approximation of P through $dP_{x_i}(x)$ leads to a better estimate of the risk. This is more formally described by Zheng et al. [316].

It should be noted that all points in the vicinity of x_i share the same label y_i in this formulation (as represented by the Dirac mass $\delta_{y_i}(y)$ in eq. 2.4) This assumption is mostly true (or should be true) for standard data augmentation techniques (e.g. Gaussian noise, crop, cutout [76], color jittering for natural images) but it is not mandatory: some works have extended the vicinal distribution $d\hat{P}_{vic}(x, y)$ to sample with different labels (e.g. Mixup [318]). Interestingly, recent work by R. Balestriero et al. [19] suggests that classical augmentations used for ImageNet (e.g. random crop) can be strongly class-dependent and shades light on the violation of this assumption for some classes.

Mental illness classification as case-study

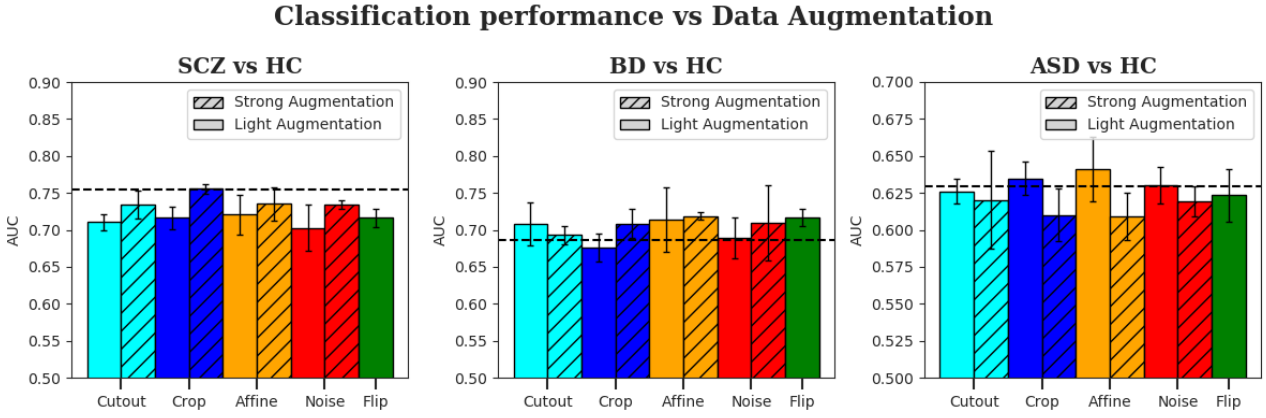


Figure 2.8: Data augmentation is strongly class-dependent and it does not result in significant improvement on clinical datasets. Applying strong augmentations can be somewhat beneficial for some classes (e.g. SCZ or BD classes with crop and affine transformation respectively) but it can lead to a constant deterioration for others (e.g ASD class). It suggests that some augmentations (e.g. strong affine transformation for ASD) create biased datasets that are not label-preserving for mental disorder classification. Baseline performance (with no augmentation) is reported with dotted lines.

Since DNN rapidly over-fit on psychiatric disorder classification tasks, we have explored several standard augmentations including geometrical transformations, random noise and cropping applied to MRI scans. Our main concern was to apply transformations i) that preserve semantic information (i.e. brain anatomical biomarkers explaining the current pathology) and ii) that were plausible (for instance artefacts or noise that can be present in real MRI scans, illustrated here by Gaussian noise). However, we acknowledge that it is difficult to know a priori what

augmentations preserve the label and only a post-hoc analysis can reveal this assumption is met.

As noted by Hernandez-Garcia [142], strong augmentations produce more biologically plausible representations compared to light augmentations (maybe because it generates examples that should be explored by DNN for good generalization on test images, exploiting domain knowledge). Interestingly, this observation may be corroborated with evidence found on current self-supervised models (SimCLR [52], BYOL [120], etc.): their exceptional representation quality depends on a very aggressive augmentation strategy. We will come back to this in the next chapter.

We have thus evaluated, for each augmentation strategy, 2 schemes (light and strong) that are described in Table 2.4. From previous analysis (see Fig. 2.2), we found that DenseNet offers good performance compared to ResNet and AlexNet on mental illness classification (especially schizophrenia and ASD). We choose this architecture to conduct the experiments. We also use the maximum number of training examples in BHB-10K, as in previous experiments, to evaluate the true utility of data augmentation in a real-world scenario.

Augmentations	**Affine**	**Crop**	**Gaussian Noise**	**Cutout**
Strong	rot(-45deg, 45deg) trans(0, 50vox) zoom(0, 0.2)	0.5*(h, w, d)	$\sigma \sim \mathcal{U}([0, 5\sigma_0])$	50% black patch
Light	rot(-5deg, 5deg) trans(0, 10vox) zoom(0, 0.1)	0.75*(h, w, d)	$\sigma \sim \mathcal{U}([0, \sigma_0])$	25% black patch

Table 2.4: Hyper-parameters cross-validated to evaluate the benefit of D.A., viewed as regularization, on final performance.

Results are plotted in Fig. 2.8. We can make several observations. First, all transformations are strongly class-dependent (e.g. flip is mostly beneficial for BD vs HC but not SCZ vs HC). Second, no augmentations stand out and it does not bring significant improvement compared to baseline and can even degrade the performance (e.g strong crop or affine transformation for ASD vs HC). This notably suggests that label-preserving assumption is not met for these transformations. Interestingly, these conclusions align well with recent findings on ImageNet [19]: some augmentations create a bias in brain imaging datasets that are not label-preserving for mental disorders (as it is the case on ImageNet for color jittering on color-dependent classes such as birds). Results obtained with affine and flip augmentations are expected since all brain images are registered with a complex non-linear pipeline [12] to the same template.

In summary, these results suggest that current augmentations crafted from human perception are not well-adapted for brain imaging tasks. Geometrical approaches based on differential geometry may be more adapted to synthesize new examples that respect the label-preserving assumption while extending the data input space (see Fig. 2.7 and [46] for a concrete application on Alzheimer's disease).

Broader analysis on phenotype prediction

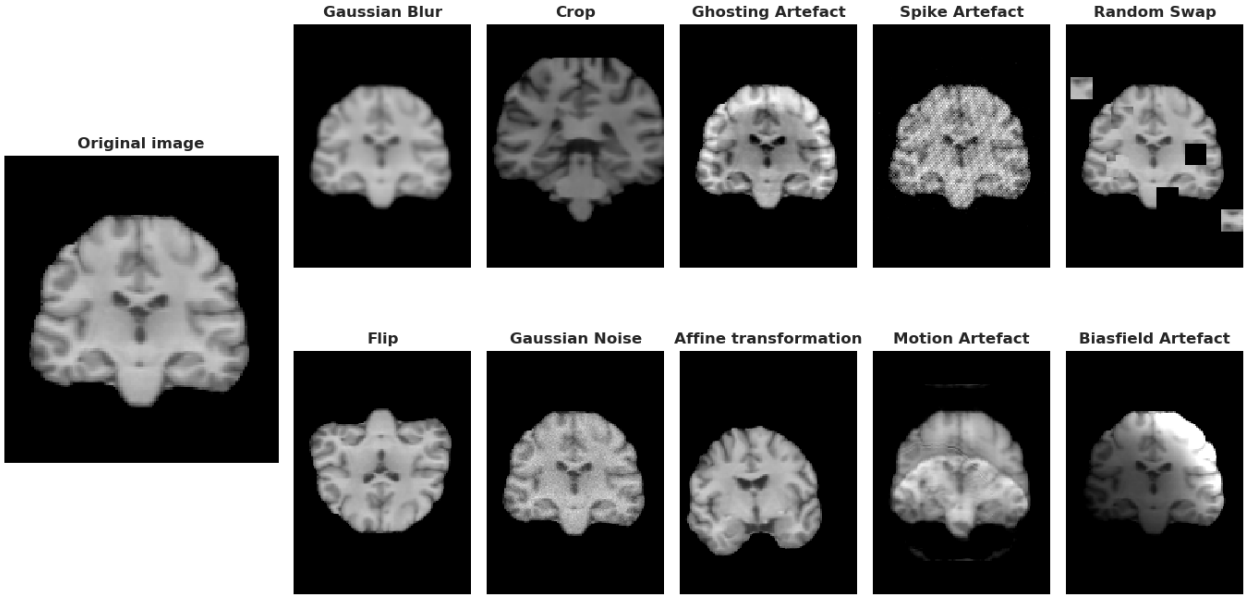


Figure 2.9: Illustration of the augmentations applied to a (quasi) raw MR image.

Since we still do not have evidence that DNN can leverage non-linear patterns for mental disorder detection, we have conducted a broader analysis of data augmentation for phenotype prediction. Previously, we showed a significant improvement of DNN compared to linear models in the large-scale data regime ($N > 9$k) for age regression. We ask: can we retrieve such improvement by adding artificially augmented images ? Does it have the same effect as adding real images ?

We conduct the experiments on BHB-10K with only $n = 500$ samples (considering the learning curves obtained Fig. 2.2) and we consider both quasi-raw and VBM images (in two distinct sets of experiments). Additionally to the previous transformations, we have evaluated Gaussian blur and artefacts that we usually observe in brain images: ghosting artefacts [322], spike artefact [322], bias-field artefact [287] and motion artefact [254] (see Fig. 2.9 Table 2.4 for more details). Finally, we also implemented swapping [49], a transformation originally introduced for self-supervision on brain images. It consists in swapping several times two patches at random location in the image. Originally, the self-supervised task consisted in decoding the original image from the latent vector given by the encoded noisy image, thus restoring back the misplaced patches from their surrounding voxels. Here, the procedure is implicit: the internal DNN representation should remove the erroneous anatomical information of the misplaced patches to correctly classify brain images.

All of these transformations, along with their hyper-parameters, are detailed in table 2.5. They have all been applied on-the-fly during training with a probability $p = 50\%$ for each input scan. The test set was never transformed and we did not apply test-time augmentation [258]

Application	Transformation	Details	Hyperparameters
Computer Vision	Flip	The images are flipped randomly along the 3 directions (axial, sagittal, coronal).	✗
	Gaussian Blur	A Gaussian filter is applied to input images with a full width at half maximum (FWHM) uniformly sampled in $[\alpha, \beta]$	FWHM $\in [0.35mm, 3.5mm]$
	Gaussian Noise	A Gaussian noise is added with a variance σ uniformly sampled in $[\alpha, \beta]$.	$\sigma \in [0.1, 1]$
	Random Crop (+Resize)	The images are cropped at a random location, reducing the input shape by $p\%$ in every direction, and resized linearly to match the input size.	Patch $p = 70\%$
	Affine	The images are randomly translated up to k voxels in every direction and rotated up to α degrees.	$k = 10$ voxels, $\alpha = 5°$
Neuroimaging	k-space Ghosting Artefact [322]	n lines in the k-space are randomly distorted to mimic the errors that may happen during the k-space line inversion step in an echo-planar imaging acquisition.	$n = 10$
	k-space Motion Artefact [254]	The image is successively randomly linearly transformed ($n_{sim}\times$, up to $\alpha°$ rotation, t voxels translation) to reproduce the head motion artefact observed during an acquisition. The 3D Fourier transforms of these images are then combined to form a single k-space, which is transformed back to the original space.	$n_{sim} = 3$, $\alpha = 40°$, $t = 10$ voxels
	k-space Spike Artefact [322]	n points with very high or low intensity are added randomly in the k-space reproducing the bad data points obtained with gradients applied at a very high duty cycle. It results in dark stripes in the original image.	$n = 10$
	Bias-Field Artefact [287]	The voxel intensities are modulated by a polynomial function (order 3, coeff. magnitude m) whose coefficients are randomly sampled. It models the artefacts in the low-frequency range produced by the inhomogeneity of the static magnetic field inside the MRI scanner.	$m \in [-0.7, 0.7]$
	Swap [49]	n pairs of patches with shape $15 \times 15 \times 15$ are randomly swapped. Originally created as a self-supervision task to learn meaningful semantic features, the network is expected to use the context around each patch in order to find its original location and internally reconstruct the image.	$n = 20$

Table 2.5: Description of the data augmentation strategies considered in our experiments. The input image always correspond to the pre-processed MR image. All the k-space artefacts have been implemented in the Python library TorchIO [219].

as the network should be already invariant to the transformations applied during training. We propose to assess the importance of each data augmentation technique separately using either VBM or quasi-raw data for age and sex prediction. To the best of our knowledge, this is the first time MRI artefacts are employed as data augmentation for such tasks. Again, we use DenseNet backbone as encoder since it performed well on all tasks (see Fig. 2.2) except for age regression on raw data (see Fig. 2.3). In that case, we trained ResNet because it was much more stable.

Finally, please note that we applied MRI artefacts only on quasi-raw images and not on VBM data since they were conceived for T1 raw images and not for gray matter density maps. Indeed, in order to apply MRI artefacts, one needs to compute the inverse Fourier transform to

map the image back to the k-space [322]. When considering VBM data, one would also need to compute the backward mapping from gray matter density to the original image and this would be computationally too demanding and prone to error.

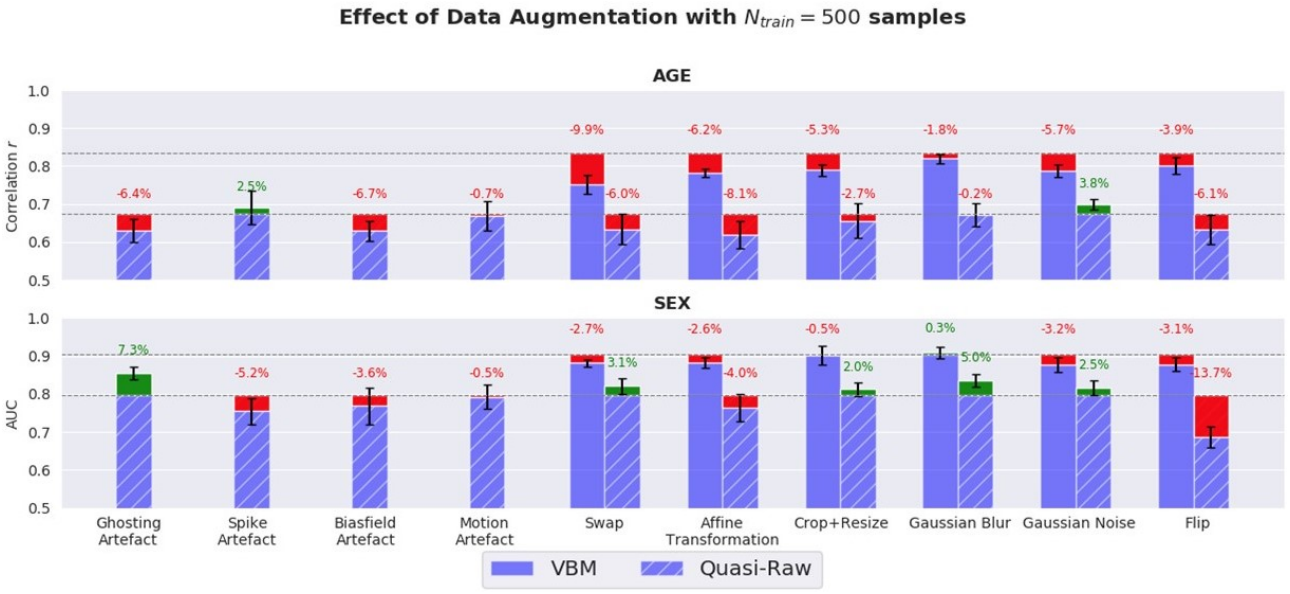


Figure 2.10: Current data augmentation (D.A) techniques are highly task- and pre-processing-dependent. It does not result in large improvement and, overall, it even degrades the performance for both VBM and quasi-raw images. The error bars are obtained using a 5-split RLT strategy using each time only one data augmentation strategy. We reported the results obtained on the external test set BSNIP ($n = 200$). The black dashed lines represent the baselines without D.A.

From Fig. 2.10, we observe that data augmentation brings little or no improvement for both VBM and quasi-raw images, retrieving the results obtained previously on clinical tasks. As opposed to previous studies [10, 67, 218], affine transformation and flip did not improve the performance on age prediction. Differently from the above-mentioned studies, our results are reported on cross-site images which may explain the differences. Also, as mentioned previously, since all images are registered to the same template (meaning all brains are well-aligned), affine transformation seems not adapted to our tasks.

Interestingly, horizontal and vertical flip degrade significantly the performance mostly for sex prediction, which may support the hemispheric asymmetry hypothesis between females and males [231] (a question still debated currently [89, 115] and with overall no clear link with behavioral data). Additionally, Gaussian blur seems to be beneficial mostly for raw data, which can be interpreted as a Gaussian smoothing effect to correct MRI reconstruction imperfections (especially since all data were resampled at 1.5mm^3 isotropic).

Again, as before, data augmentation is both task- and pre-processing-dependent and it does not necessarily result in large improvement neither for regression nor classification tasks. For an easy task (sex prediction with $AUC \geq 0.9$) it significantly improves the performance only with quasi-raw data (*i.e*, with ghosting artefact or Gaussian blur). This mitigates the usefulness of current data augmentation techniques on brain MRI, especially when all images have been

aligned to the same template and re-sampled to the same spatial resolution. Even with the minimal pre-processing (i.e., quasi-raw), there is no clear improvement with the standard D.A (affine transformation, Gaussian blur, etc.). Furthermore, we also showed that adding MRI artefacts into the data augmentation strategy brings overall no improvement and it actually worsen the results most of the time (except for ghosting artefact and spike artefact for sex and age prediction respectively).

Perspectives. Our work contains several limitations. First, we did not cross-validated all possible hyper-parameter values for each transformation (as we have a limited computational budget). Second, our study was performed in the supervised setting with binary cross-entropy loss for classification and ℓ_1 loss for regression. This design seems rather standard however it may imply strong assumptions about the optimization landscape that may explain our results on supervised tasks. For instance, we know that SupCon, another supervised loss for classification introduced for contrastive learning (see next chapter), is more robust to noisy input than cross-entropy [118]. The underlying reason is still debated but it could indicate that cross-entropy is not the best choice for neuroimaging data. Finally, this work does not imply that data augmentation is not a good tool for neuroimaging but rather that most augmentations crafted from human perception on natural images does not translate well for brain imaging. In fact, very recently a new geometry-aware Variational AutoEncoder (VAE) has been invented [46] to generate synthetic images of Alzheimer Disease (AD) brains and healthy controls (HC). By augmenting artificially the training set, it shows a significant improvement on small-scale datasets for discriminating AD vs HC.

2.4.2 Data harmonization as data-based debiasing strategy

As reported in several multi-site studies [112, 293], the high heterogeneity between scanners and acquisition protocols has led ML models to under-perform on data coming from other sites than the ones used during training (a.k.a, domain gap on out-of-distribution samples). We notably confirmed and extended these results for a wider range of brain imaging tasks in the previous section (e.g. Fig. 2.2). As we saw, this heterogeneity indeed leads to a consistent performance drop for DL and SML models between internal and external test sets.

Schulz [249, 250] hypothesized that current noise in brain imaging linearizes the decision boundary for phenotype prediction and, here, for psychiatric disorders classification. We provided evidence supporting this hypothesis and we notably showed how site-related information (viewed as noise) were well-preserved in DNN representations (see Table 2.3). We may wonder: can we remove this noise from the datasets ? By doing so, can we improve DNN representations over linear models ?

To answer these questions, we used two SOTA harmonization methods to remove site information, viewed as a confounding variable: ComBat [101, 161] and Linear Adjusted Regression. These two methods directly harmonize the data without changing the model (as opposed to re-

cent methods [82] acting on DL representations), allowing a fair comparison between SML and DL methods. Importantly, both ComBat and Linear Adjusted Regression need image statistics on all sites to remove site information. However, in our case, only training and internal test set contain the same sites so we only residualized these two sets, leaving the external test unchanged. We describe briefly this two models before showing the actual results.

ComBat and linear adjusted regression models

Linear Adjusted Regression. It is a simple linear harmonization method that tries to preserve biological variability from the data, while removing non-biological effect (such as site effect). The model itself can be expressed as [293]:

$$\begin{cases} Y_{ijf} = \alpha_f + \gamma_{if} + {\beta_\mathbf{f}}^T \mathbf{k}_j + \epsilon_{ijf} \\ \mathcal{Y}_{ijf} = Y_{ijf} - \hat{\gamma}_{if} \end{cases}$$

where Y_{ijf} is the voxel value for site i, subject j, voxel f; α_f is an average measure for voxel f, γ_{if} is the site effect, $\mathbf{k}_j$ is the vector of biological variables we want to keep for subject j (i.e age, sex and diagnosis eventually) and β_f are parameters estimated by linear regression. $\mathcal{Y}_{ijf}$ is the residualized voxel value, where $\hat{\gamma}_{if}$ is the estimated site effect. The parameters γ_{if} and β_f are estimated during training.

ComBat [101]. Differently from linear adjusted regression, it adds a multiplicative non-linear effect δ_{if} on the residual noise ϵ_{ijf} which brings to a different residualization scheme that also requires the biological variables $\mathbf{k}_j$:

$$\begin{cases} Y_{ijf} = \alpha_f + \gamma_{if} + {\beta_\mathbf{f}}^T \mathbf{k}_j + \delta_{if}\epsilon_{ijf} \\ \mathcal{Y}_{ijf} = \frac{Y_{ijf} - \hat{\alpha}_f - \hat{\beta}_f^T \mathbf{k}_j - \hat{\gamma}_{if}}{\hat{\delta}_{if}} + \hat{\alpha_f} + \hat{\beta}_f^T \mathbf{k}_j \end{cases}$$

Biased results with ComBat. An attentive reader may have noticed that, unlike linear adjusted regression, ComBat needs the biological variables $\mathbf{k}_j$ to perform residualization (i.e. compute $\mathcal{Y}_{ijf}$). In our case, this is a clear "data leakage" (described as "Late split" in [304]) since we aim to predict these biological variables (age, sex, diagnosis) on an independent test set where they are theoretically unknown. Put differently, ComBat model introduces a bias in the (testing) data during residualization that may also lead to biased (over-optimistic) results in ML studies targeting biological variables. To our knowledge, this issue is not reported in the current literature (e.g. [221]).

(Unbiased) External test residualization. Both ComBat and Linear Adj. Regression model require to have access to all imaging sites to estimate their parameters. In our experimental design, only internal test has overlapping sites with training so we can only perform residualization on this set, leaving external test unchanged. This way, we also avoid the bias

introduced by ComBat mentioned above. Formally, we propose to set $\delta_{if} = 1$ and $\gamma_{if} = 0$ for all unknown test sites i in both linear adjusted regression and ComBat. We acknowledge this is not ideal and other DL-based [82, 278] solutions are starting to emerge to remove site-effect but there is still no consensus and most of the current studies still use ComBat or Linear Adjusted Regression [20, 228].

DL vs SML after data harmonization

From Tab. 2.6, we observe that residualization does not bring improvement for DL models while it marginally improves performance for linear models when trained with $N_{train} = 9253$ on age prediction (-0.48 MAE for Ridge Regression on internal test). However, the difference is more pronounced on psychiatric datasets with a gain of $1-3\%$ AUC overall on the three tasks with SML models (linear and kernel-SVM). We do observe degradation in performance with DL models on both internal and external test sets, indicating that current residualization methods fail to preserve non-linear biological variability that was extracted by DL models. We performed additional experiments on DenseNet and ResNet clearly supporting these conclusions in Table 2.6.

Task	Model	Internal Test			External Test		
		Linear Adj. Res.	ComBat	No Res.	Linear Adj. Res.	ComBat	No Res.
Age ↓ $N_{train} = 9253$	AlexNet	$2.79_{\pm 0.07}$	$2.98_{\pm 0.06}$	$\mathbf{2.36_{\pm 0.04}}$	$4.59_{\pm 0.08}$	$6.92_{\pm 1.03}$	$\mathbf{3.43_{\pm 0.02}}$
	rbf-SVM	$3.34_{\pm 0.00}$	$3.67_{\pm 0.00}$	$\mathbf{3.21_{\pm 0.00}}$	$4.59_{\pm 0.00}$	$5.74_{\pm 0.00}$	$\mathbf{4.27_{\pm 0.00}}$
	Ridge	$\mathbf{3.08_{\pm 0.00}}$	$3.33_{\pm 0.00}$	$3.56_{\pm 0.00}$	$4.93_{\pm 0.00}$	$4.39_{\pm 0.00}$	$\mathbf{4.21_{\pm 0.00}}$
	ElasticNet	$\mathbf{3.14_{\pm 0.00}}$	$3.21_{\pm 0.02}$	$3.31_{\pm 0.00}$	$4.62_{\pm 0.00}$	$4.38_{\pm 0.03}$	$\mathbf{4.25_{\pm 0.00}}$
Sex ↑ $N_{train} = 9253$	AlexNet	$93.88_{\pm 0.64}$	$95.24_{\pm 0.55}$	$\mathbf{96.13_{\pm 0.42}}$	$94.54_{\pm 0.34}$	$95.58_{\pm 0.65}$	$\mathbf{97.91_{\pm 0.15}}$
	rbf-SVM	$\mathbf{96.09_{\pm 0.00}}$	$95.86_{\pm 0.00}$	$95.16_{\pm 0.00}$	$97.88_{\pm 0.00}$	$98.03_{\pm 0.00}$	$\mathbf{97.28_{\pm 0.00}}$
	Logistic	$95.88_{\pm 0.04}$	$95.63_{\pm 0.03}$	$\mathbf{95.95_{\pm 0.04}}$	$98.26_{\pm 0.00}$	$98.23_{\pm 0.03}$	$\mathbf{98.32_{\pm 0.00}}$
	ElasticNet	$95.09_{\pm 0.05}$	$94.83_{\pm 0.01}$	$\mathbf{95.23_{\pm 0.01}}$	$\mathbf{98.04_{\pm 0.04}}$	$97.95_{\pm 0.65}$	$97.93_{\pm 0.05}$
SCZ vs HC ↑ $N_{train} = 933$	AlexNet	$71.53_{\pm 0.71}$	$\mathbf{82.35_{\pm 1.45}}$	$79.13_{\pm 0.96}$	$68.50_{\pm 0.90}$	$\mathbf{74.14_{\pm 1.13}}$	$72.07_{\pm 0.95}$
	rbf-SVM	$\mathbf{83.55_{\pm 0.00}}$	$82.06_{\pm 00}$	$82.06_{\pm 0.00}$	$\mathbf{76.39_{\pm 0.00}}$	$72.88_{\pm 0.00}$	$72.88_{\pm 0.95}$
	Logistic	$\mathbf{85.31_{\pm 0.07}}$	$84.25_{\pm 0.02}$	$84.03_{\pm 0.03}$	$\mathbf{76.45_{\pm 0.15}}$	$73.76_{\pm 0.46}$	$73.60_{\pm 0.00}$
	ElasticNet	$\mathbf{88.81_{\pm 1.03}}$	$86.96_{\pm 0.82}$	$85.98_{\pm 1.9}$	$78.98_{\pm 0.98}$	$\mathbf{79.02_{\pm 1.08}}$	$76.42_{\pm 1.68}$
BD vs HC ↑ $N_{train} = 832$	AlexNet	$62.41_{\pm 3.03}$	$66.77_{\pm 5.44}$	$\mathbf{74.16_{\pm 3.25}}$	$61.67_{\pm 1.26}$	$65.58_{\pm 1.73}$	$\mathbf{72.46_{\pm 2.74}}$
	rbf-SVM	$\mathbf{75.00_{\pm 0.00}}$	$70.92_{\pm 0.00}$	$73.63_{\pm 0.00}$	$\mathbf{67.74_{\pm 0.00}}$	$63.36_{\pm 0.00}$	$63.92_{\pm 0.00}$
	Logistic	$\mathbf{74.07_{\pm 0.09}}$	$73.17_{\pm 0.38}$	$72.96_{\pm 0.25}$	$69.54_{\pm 0.33}$	$69.36_{\pm 0.28}$	$\mathbf{70.12_{\pm 0.26}}$
	ElasticNet	$71.19_{\pm 2.29}$	$72.27_{\pm 1.60}$	$\mathbf{73.85_{\pm 0.28}}$	$\mathbf{70.33_{\pm 2.47}}$	$68.14_{\pm 0.93}$	$70.26_{\pm 1.75}$
ASD vs HC ↑ $N_{train} = 1526$	AlexNet	$59.06_{\pm 1.96}$	$58.55_{\pm 1.34}$	$\mathbf{62.07_{\pm 1.77}}$	$54.25_{\pm 2.06}$	$60.51_{\pm 1.09}$	$\mathbf{62.46_{\pm 1.21}}$
	rbf-SVM	$66.78_{\pm 0.00}$	$64.64_{\pm 0.00}$	$\mathbf{66.84_{\pm 0.00}}$	$59.10_{\pm 0.00}$	$58.94_{\pm 0.00}$	$\mathbf{60.28_{\pm 0.00}}$
	Logistic	$\mathbf{64.71_{\pm 0.22}}$	$63.11_{\pm 0.09}$	$63.40_{\pm 0.18}$	$\mathbf{63.98_{\pm 0.15}}$	$61.98_{\pm 0.30}$	$61.85_{\pm 0.05}$
	ElasticNet	$\mathbf{63.30_{\pm 4.78}}$	$60.30_{\pm 3.76}$	$60.62_{\pm 2.63}$	$57.98_{\pm 4.71}$	$\mathbf{60.21_{\pm 3.19}}$	$54.96_{\pm 4.94}$

Table 2.6: DL vs SML performance on residualized data. Current residualization techniques are particularly well-suited for linear models (consistent improvement, +1-3% AUC, of ℓ_2-penalized linear regression on all clinical tasks). Kernel-SVM also highly benefit from residualization (+4% AUC on SCZ vs HC and BD vs HC on external test). Interstingly, more consistent improvements (between 1% and 3% AUC) appear with less training samples ($N_{train} < 2000$) on diagnosis classification tasks with SML. On the contrary, DL models under-perform for all tasks on these data, showing no improvement w.r.t linear models (see also Fig. 2.7 for more results with DenseNet121 and ResNet18). AlexNet is reported as representative of CNN models. All models are trained 3 times with different random initialization and standard deviation is reported. AUC is reported for binary classification tasks, while MAE is reported for age prediction.

DL performance on residualized data. To confirm the previous results obtained with AlexNet architecture, we also trained DenseNet and ResNet on the same data residualized with linear adjusted regression (protecting age, sex and diagnosis). In Table 2.7, we observe a constant decrease in performance when performing residualization.

Task	Model	Internal Test		External Test	
		Linear Adj. Res.	No Res.	Linear Adj. Res.	No Res.
Age ↓ $N_{train} = 9253$	AlexNet	$2.79_{\pm 0.07}$	$\mathbf{2.36_{\pm 0.04}}$	$4.59_{\pm 0.00}$	$\mathbf{3.43_{\pm 0.02}}$
	DenseNet	$2.75_{\pm 0.06}$	$\mathbf{2.58_{\pm 0.09}}$	$4.24_{\pm 0.01}$	$\mathbf{3.53_{\pm 0.07}}$
	ResNet18	$2.75_{\pm 0.06}$	$\mathbf{2.49_{\pm 0.08}}$	$3.76_{\pm 0.03}$	$\mathbf{3.49_{\pm 0.08}}$
Sex ↑ $N_{train} = 9253$	AlexNet	$93.88_{\pm 0.64}$	$\mathbf{96.13_{\pm 0.42}}$	$94.54_{\pm 0.34}$	$\mathbf{97.91_{\pm 0.15}}$
	DenseNet	$94.55_{\pm 0.03}$	$\mathbf{96.57_{\pm 0.25}}$	$95.48_{\pm 0.16}$	$\mathbf{98.47_{\pm 0.11}}$
	ResNet18	$95.46_{\pm 0.40}$	$\mathbf{96.33_{\pm 0.34}}$	$96.72_{\pm 0.40}$	$\mathbf{98.39_{\pm 0.26}}$
SCZ vs HC ↑ $N_{train} = 933$	AlexNet	$71.53_{\pm 0.71}$	$\mathbf{79.13_{\pm 0.96}}$	$68.50_{\pm 0.90}$	$\mathbf{72.07_{\pm 0.95}}$
	DenseNet	$73.09_{\pm 1.32}$	$\mathbf{85.27_{\pm 1.60}}$	$63.34_{\pm 1.10}$	$\mathbf{75.52_{\pm 0.12}}$
	ResNet18	$78.12_{\pm 1.82}$	$\mathbf{80.93_{\pm 3.16}}$	$73.07_{\pm 2.15}$	$\mathbf{74.31_{\pm 0.12}}$
BD vs HC ↑ $N_{train} = 832$	AlexNet	$62.41_{\pm 3.03}$	$\mathbf{74.16_{\pm 3.25}}$	$61.67_{\pm 1.26}$	$\mathbf{65.49_{\pm 0.91}}$
	DenseNet	$62.91_{\pm 2.20}$	$\mathbf{76.49_{\pm 2.16}}$	$61.70_{\pm 3.50}$	$\mathbf{68.57_{\pm 4.72}}$
	ResNet18	$62.59_{\pm 0.85}$	$\mathbf{68.63_{\pm 3.82}}$	$67.31_{\pm 1.09}$	$\mathbf{69.33_{\pm 0.60}}$
ASD vs HC ↑ $N_{train} = 1526$	AlexNet	$59.06_{\pm 1.96}$	$\mathbf{62.07_{\pm 1.77}}$	$54.25_{\pm 2.06}$	$\mathbf{62.46_{\pm 1.21}}$
	DenseNet	$61.33_{\pm 3.25}$	$\mathbf{65.74_{\pm 1.47}}$	$54.70_{\pm 2.07}$	$\mathbf{62.93_{\pm 2.40}}$
	ResNet18	$\mathbf{59.02_{\pm 2.37}}$	$58.52_{\pm 3.25}$	$58.64_{\pm 1.66}$	$\mathbf{62.09_{\pm 1.75}}$

Table 2.7: DL performance on VBM data residualized with linear adjusted residualization (adjusted on age, sex, site and eventually diagnosis). DL performance on VBM data not residualized is indicated for comparison purposes. Linear residualization hurts performance for all models and tasks, indicating that it removes discriminative features used by DL models.

2.5 Know what you don't know helps: deep uncertainty estimation in supervised learning

Previously, we have seen that, contrary to current expectations, DNN models are not able to generalize better than (regularized) linear models on anatomical brain imaging, at least for mental disorder diagnosis. They tend to rapidly over-fit on noisy features (e.g. acquisition scanner for multi-site datasets), and current data-based harmonization methods do not bring a satisfactory solution for removing this noise.

One known issue when training DNN with cross-entropy loss is their over-confidence in their prediction. Concretely, as the optimization goes, the network starts to become over-confident in all its prediction (not only samples inside the training distribution but also on out-of-distribution samples), even when its prediction is wrong. We have illustrated this on a toy example in Fig. 2.11 (left). Notably, in 2016, it has been shown [122] that modern DNN architectures (e.g. ResNet) are far more over-confident than a decade ago with simple LeCun architecture. This has been (partly) attributed to the current over-parametrization of DNN

(e.g. ResNet110 with 110 layers or DenseNet121 with 121 layers vs LeCun with 5 layers) that has led to a serious degradation in calibration, although the accuracy of such networks were also drastically increased. The fundamental reason why DNN are so good at generalizing even in the heavily over-parameterized regime is still poorly understood. Nevertheless, having poorly calibrated classifiers for critical applications such as computer-aided diagnosis is a serious issue since we cannot reasonably trust such classifiers. In a real-world scenario where an AI system helps an expert to screen MRI scanners for, let's say, the prognosis of First-Episode Psychosis (FEP) within a year, having an over-confident system can strongly bias the expert's prognosis. This could mislead its judgement by asserting a strong statement with high confidence and it is clearly not acceptable.

Modelling uncertainty inside current deep networks is fairly recent (see for instance Gal et al. [105]) and is mostly based on Bayesian theory. In our case, we saw Section 2.3 that deeper models with more parameters (e.g. DenseNet121 vs AlexNet) did not result in a significant gain in performance for the current datasets size. In this section, we ask: by improving uncertainty estimation in highly over-parameterized networks, can we improve performance to outperform linear models on brain imaging ?

We answer to this question by studying two main paradigms to model uncertainty in DNN predictions: Monte-Carlo dropout (MC-Dropout) [106] and Deep Ensemble learning [183]. Both methods learns an approximation of $p(y|x)$, distribution of the target label y given the input data x, by modeling both *aleatoric uncertainty* (related to irreducible noise in the data) and *epistemic uncertainty* (associated to uncertainty in model's parameters, which is often disregarded during training). In the next sections, we first give a more formal definition of epistemic and aleatoric uncertainty through Bayesian DNN theory before introducing MC-Dropout and Deep Ensemble learning models. We then present and discuss the results.

2.5.1 Aleatoric and epistemic uncertainty in DNN

We consider a supervised problem where we want to predict a target $y \in \mathcal{Y}$ (continuous $\mathcal{Y} = \mathbb{R}^p$ or categorical $\mathcal{Y} = [1..C]$ with C classes) from an input image $x \in \mathcal{X}$. A DNN is defined as a mapping $f_\theta : \mathcal{X} \mapsto \mathcal{U}$ from an image $x \in \mathcal{X}$ to an output $f_\theta(x) \in \mathcal{U}$, parametrized by $\theta \in \mathbb{R}^l$. This mapping f_θ models a target distribution $p(y|x, \theta)$ that integrates the *aleatoric uncertainty*. It is intrinsic to the input data x (e.g. artefacts in MRI viewed as noise) so it is irreducible (we have to deal with the noisy data we have).

Classification. In a classification setting, $f_\theta(x)$ gives the logit scores and the underlying distribution is, for any $i \in [1..C]$:

$$p(y = i|x, \theta) = \text{softmax}(f_\theta(x))[i] = \frac{e^{f_\theta(x)[i]}}{\sum_{k=1}^{C} e^{f_\theta(x)[k]}} \tag{2.6}$$

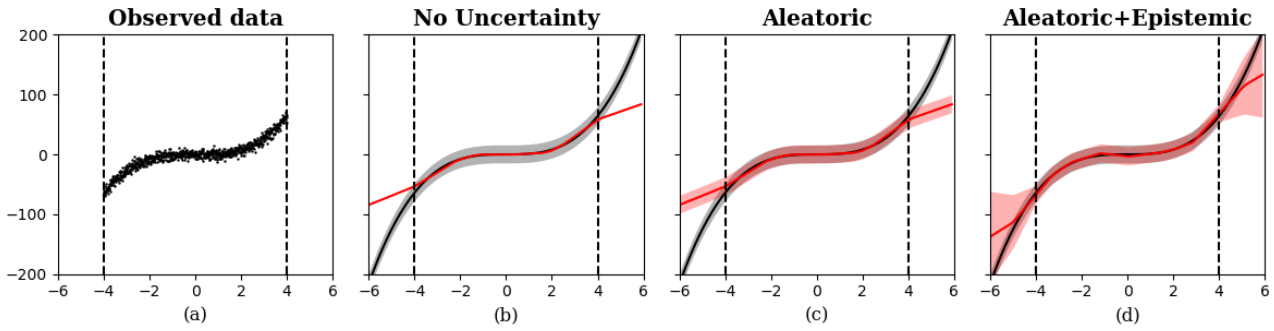


Figure 2.11: Illustration of deep uncertainty estimation on a toy regression task (adapted from [123]). True distribution $p(y|x)$ is a Gaussian with mean in solid black line and variance in shaded gray. Predictive mean and variance are given with solid red line and shaded red area. (a) Training data $\{(x_i, y_i)\}_{i=1}^{800}$ generated from a Gaussian distribution $p(y|x) = \mathcal{N}(\mu(x), \sigma^2)$ with $\mu(x) = x^3$ and $\sigma = 5$. (b) DNN trained to predict y from x with ℓ_2 loss. (c) DNN trained to optimize likelihood of $p(y|x, \theta) = \mathcal{N}(\mu_\theta(x), \sigma_\theta^2)$, capturing only aleatoric uncertainty (d) DNN trained to minimize $\hat{p}(y|x, \mathcal{D})$ (eq. 2.8), Bayesian approximation of $p(y|x, \mathcal{D})$ capturing both aleatoric and epistemic uncertainty. We hypothesize that a better approximation of $p(y|x, \mathcal{D})$ improves i) calibration and ii) performance.

For a training set with n examples $\mathcal{D} = (X, Y) = \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} p(x, y)$, the likelihood is $p(Y|X, \theta) = \prod_{i=1}^n p(y_i|x_i, \theta)$ and the negative log-likekihood (NLL) can be expressed as $\mathcal{L}_{NLL} = -\sum_{i=1}^n p(y_i|x_i, \theta)$ which is also called cross-entropy loss. Thus, the optimal $\hat{\theta}_{MV}$ minimizing this loss is called the maximum log-likehood estimator.

Regression. For a regression task, $f_\theta(x)$ usually gives a single value $\hat{y}$ and it does not translate directly into a distribution $p(y|x, \theta)$. As a result, the DNN does not model directly an aleatoric uncertainty. One usual solution [169] is to model $p(y|x, \theta)$ as a Gaussian distribution $\mathcal{N}(\mu_\theta(x), \sigma_\theta^2(x))$ whose parameters are given by $f_\theta(x) = [\mu_\theta(x), \sigma_\theta^2(x)]$. If we assume homoscedasticity (i.e. a variance σ_θ^2 independent of x) then minimizing the NLL of $p(Y|X, \theta)$ is equivalent to minimizing the ℓ_2 loss with an extra parameter σ_θ^2 to learn[4].

Epistemic uncertainty. Previous models only take into account aleatoric uncertainty, accounting for noisy data. The *epistemic uncertainty*, associated to model's parameters θ is never included. That's where Bayesian inference comes into play. We use the posterior distribution $p(\theta|\mathcal{D})$ to define the predictive posterior distribution (see Appendix A.1):

$$p(y|x, \mathcal{D}) = \int p(y|x, \theta) p(\theta|\mathcal{D}) d\theta \tag{2.7}$$

This distribution integrates both i) aleatoric uncertainty (through $p(y|x, \theta)$ and ii) epistemic uncertainty (through $p(\theta|\mathcal{D})$). In practice, to compute this integral, one would need to perform Monte-Carlo sampling by using T samples $\theta^{(i)} \sim p(\theta|\mathcal{D})$. However, $p(\theta|\mathcal{D}) \propto p(Y|X, \theta)p(\theta)$ is not accessible and we must use an approximation $q(\theta) \approx p(\theta|\mathcal{D})$. The approximate predictive

[4]It becomes clear from equality $\mathcal{L}_{NLL} = \frac{1}{2}\sum_{i=1}^n - \log(2\pi\sigma_\theta^2) + \frac{1}{\sigma_\theta^2}||y_i - \mu_\theta(x_i)||^2$

posterior distribution can be expressed as:

$$\hat{p}(y|x,\mathcal{D}) = \frac{1}{T}\sum_{i=1}^{T} p(y|x,\theta^{(i)}) \quad \theta^{(i)} \sim q(\theta) \tag{2.8}$$

The main question is now, what approximation $q(\theta)$ can we use to accurately approximate the distribution $p(\theta|\mathcal{D})$?

2.5.2 Deep ensemble learning

In [183], authors introduced deep ensemble learning as a simple method to sample according to an approximation of $p(\theta|\mathcal{D})$. It consists in training independently T identical DNN with different initialization $(\theta_0^{(t)})_{t\in[1,..T]}$ and shuffling the data during the stochastic gradient descent optimization step. At the end of the optimization, this gives T models $f_{\theta^{(t)}}$ where each model's weights $\theta^{(t)} \sim q(\theta) \approx p(\theta|\mathcal{D})$. The hope is that $q(\theta)$ provides a good approximation of $p(\theta|\mathcal{D}) \propto p(Y|X,\theta)p(\theta)$. $p(\theta|\mathcal{D})$ is highly multi-modal because of $p(Y|X,\theta)$ [123]. So intuitively the main hypothesis is that local minima $\theta^{(t)}$ obtained by optimizing the likelihood $p(Y|X,\theta)$ will capture the main modes of $p(\theta|\mathcal{D})$ (see Fig. 2.13).

2.5.3 MC-Dropout

MC-Dropout has been introduced by Gal et al.[105, 106] as a rough approximation of $p(\theta|\mathcal{D})$ using a Bernouilli prior distribution $\mathcal{B}(p)$. It has been successfully applied in the medical imaging field to diabetic retinopathy diagnosis [95, 191]. Concretely, for each (variational) parameter θ_i in the DNN, we define the distribution $q(\theta_i) = \theta_i \cdot z_i$ where $z_i \sim \mathcal{B}(p_i)$ with a probability p_i. This way, $q(\theta) = \prod_{i=1}^{l} q(\theta_i)$ is a highly multi-modal distribution with high correlations between the weights $\theta = (\theta_i)_{i\in[1..l]}$.

In addition to its simplicity, this model gives a readable interpretation of dropout technique [265] in current DNN. Previously, dropout was used mainly as a regularization technique to limit over-fitting during training. Here, Gal et al. showed that adding dropout corresponds to a Monte-Carlo sampling over a variational distribution $q(\theta)$ to approximate the predictive posterior distribution $p(y|x,\mathcal{D})$. This notably implies that it can be used both during training *and* test to compute $\hat{p}(y|x,\mathcal{D})$ as in eq. 2.8 and to integrate aleatoric *and* epistemic uncertainties. In practice, sampling $\theta^{(i)} \sim q(\theta)$ and computing $p(y|x,\theta^{(i)})$ corresponds to a single feed-forward pass in the DNN with dropout activated.

As the reader may have notice, MC-Dropout introduces a hyper-parameter p_i for each DNN parameter θ_i. For current networks with several hundred million parameters, it is clearly not doable to cross-validate all (p_i). Two solutions are available: either all p_i are set to the same probability p and it requires the cross-validation of a single hyper-parameter or these "hyper-parameters" can be learnt during optimization. The main difficulty to optimize $(p_i)_{i\in[1..l]}$ is the non-differentiability of the binary masks $(z_i) \sim \mathcal{B}(p_i)$ which prohibits gradient-descent

algorithm. One workaround proposed by Gal et al. [107] is to relax the Bernouilli distribution with a continuous Concrete distribution. This technique allows to perform gradient-descent on all parameters $\{\theta, (p_i)\}$ during optimization of the loss function; it avoids the cross-validation of hyper-parameters $(p_i)_{i\in[1..l]}$ and it is scalable to highly over-parametrized networks. We use this technique for this study.

2.5.4 Evaluation metrics

We recall that our main motivation in this study is to show that, by improving uncertainty estimation in over-parametrized DNN, we can improve performance and provide more reliable classifier/regressor. In practice, to evaluate model's uncertainty quality we rely on the notion of calibration. Intuitively, a well calibrated classifier should give a probability for a given class equals to its occurrence's probability (see below). A mis-calibrated model indicates that it makes under or over-confident predictions. It is usually measured by the Expected Calibration Error (ECE) that gives the confidence error between a perfectly calibrated model and the model at hand. This metric can be extended to regression problems with the Area Under Calibration Error (AUCE) score as introduced in [123].

Calibration for classification

Let's assume that a DNN outputs a class prediction y as well as a confidence estimate $\hat{p}$ (usually the maximum probability after softmax) for a given x. We want to evaluate this estimation of confidence through a "calibration curve". Intuitively, if a network outputs a class $y = 0$ with a confidence level $\hat{p} = 0.6$, then we would like that, over 100 predictions of samples belonging to class 0, 60 are correctly classified. More formally, we introduce a notion of accuracy for a given confidence level p as $p(\mathbf{y} = y|\hat{\mathbf{p}} = p)$. A perfectly calibrated model should always verify:

$$\forall p \in [0, 1], \forall y \in [1..K], p(\mathbf{y} = y|\hat{\mathbf{p}} = p) = p$$

in a classification problem with K classes. In practice, this accuracy has to be estimated for various confidence levels p and given a class k. To do so, we discretize uniformly the predicted confidence levels $\hat{p} = (\hat{p}_i)$ into L bins $I_l = [\frac{l-1}{L}, \frac{l}{L})$ and compute the accuracy of the predictions over each bin $\hat{P}_l = \{i|\frac{l-1}{L} \leq \hat{p}_i < \frac{l}{L}\}$ by:

$$acc(\hat{P}_l) = \frac{1}{|\hat{P}_l|} \sum_{i\in\hat{P}_l} \mathbb{1}_{y_i=k}$$

The estimation of the confidence level associated to the bin l, independent from class k, is then:

$$conf(\hat{P}_l) = \frac{1}{|\hat{P}_l|} \sum_{i\in\hat{P}_l} \hat{p}_i$$

In a perfectly calibrated model, we expect $\forall l \in [1..L], acc(\hat{P}_l) = conf(\hat{P}_l)$. One visual way to check the model calibration is to plot the accuracy function of confidence, the ideal case being $acc = conf$. A usual statistic derived from this calibration curve is called Expected Calibration Error (ECE) and it is defined as [122]:

$$ECE = \sum_{l=1}^{L} \frac{|\hat{P}_l|}{n} \Big(acc(\hat{P}_l) - conf(\hat{P}_l) \Big)$$

where n is the total number of samples. We systematically used this metric to measure calibration on classification problems (e.g sex prediction and mental disorder classification).

Calibration for regression

We can extend the ECE metric to the regression case, as detailed in [123]. Briefly, assuming that the model outputs a mean μ and variance σ^2 of a Gaussian distribution for a given x, we can build a confidence interval $CI(p) = [\mu - \Phi^{-1}\left(\frac{p+1}{2}\right)\sigma, \mu + \Phi^{-1}\left(\frac{p+1}{2}\right)\sigma]$ associated to a confidence level p (where Φ is the Cumulative Distribution Function, CDF, of $\mathcal{N}(0,1)$). We can compute the proportion $\hat{p}$ of true target points $y \in \mathbb{R}$ that lie in $CI(p)$, for all $p \in [0,1]$. From this, similarly to ECE, we can deduce the Area Under the Calibration Error (AUCE) of $|\hat{p} - p|$.

2.5.5 Results

We have evaluated Deep Ensemble learning and MC-Dropout in the context of brain imaging prediction tasks. We have chosen 3 representative tasks: age regression, sex classification (easy) and schizophrenia detection (hard); and a limited training size ($N_{train} = 500$ considering the learning curves observed Fig. 2.2-performance is still improving for all tasks in this regime). We performed two sets of experiments: one with a very deep model (DenseNet121 with 121 layers and 11M parameters) and the other with its tiny version (tiny-DenseNet with 73 layers and 1.8M parameters, see Appendix A.2). By doing so, we can: i) check our hypothesis on several networks, ii) verify if deeper networks leads to a degradation in calibration on brain imaging tasks, as observed on common vision datasets by Guo et al. [122].

Experiments on DenseNet121

We first show the results with DenseNet121 in Fig. 2.12. Confidence bars are obtained by repeating each experiment 5 times and the standard deviation is reported. We observe a constant improvement for all metrics (both performance measured by AUC and calibration measured by ECE/AUCE) as the number of samples T used to estimate $p(y|x, \mathcal{D})$ increases (see eq. 2.8). It suggests that both MC-Dropout and Deep Ensemble provide a suitable variational approximation $q(\theta)$ of $p(\theta|\mathcal{D})$, while they rely on very different assumptions. For Deep Ensemble, it seems that training a heavily over-parametrized network such as DenseNet121 from different random

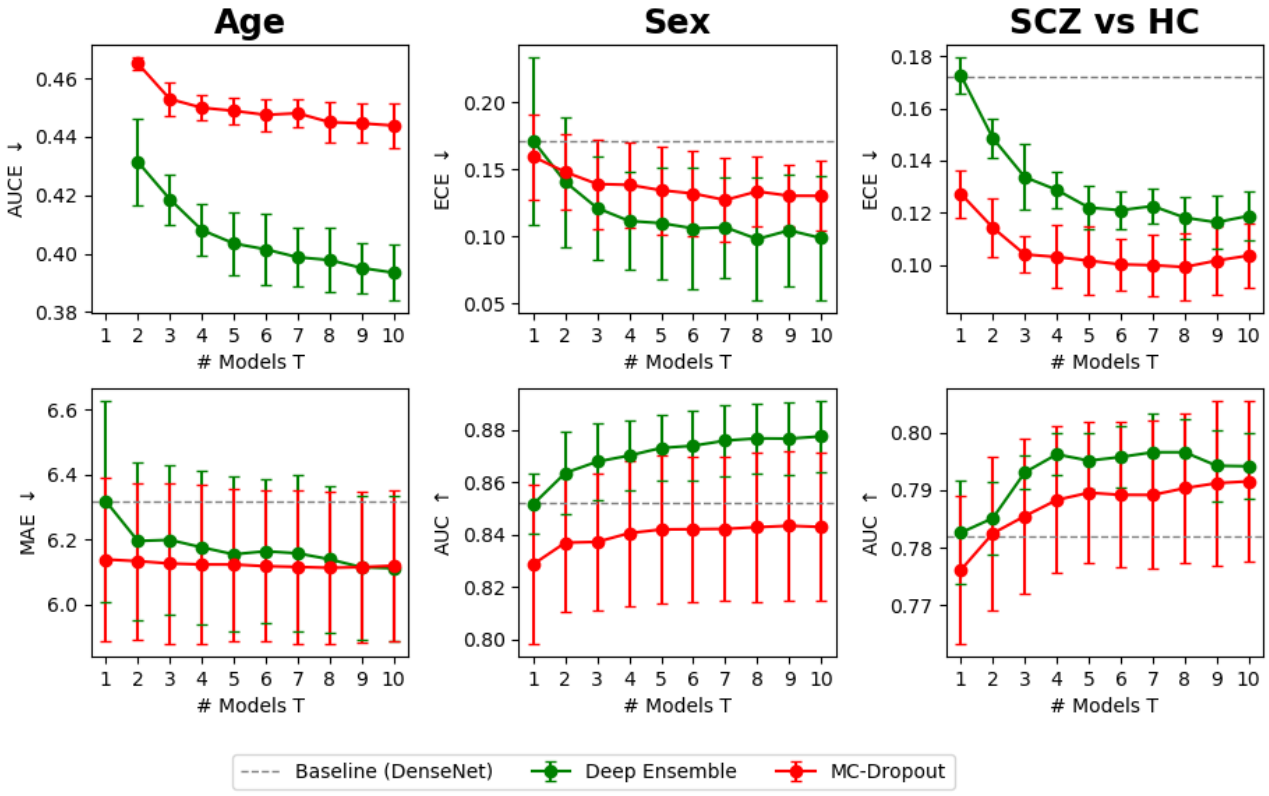


Figure 2.12: Predictive uncertainty quality (top) and performance (bottom) for Deep Ensemble model vs MC-Dropout on a regression task (age) and two classification tasks (sex and SCZ vs HC) using brain MRI. As the number of samples T used to estimate the posterior distribution $p(y|x, \mathcal{D})$ increases, both performance and uncertainty quality improve, using either Deep Ensemble or MC-Dropout. This confirms our hypothesis that having a better calibrated model (with good predictive uncertainty) leads to improved performance. Overall, Deep Ensemble offers better or equal performance for all tasks with $T = 10$ compared to MC-Dropout while being equally or better calibrated.

initialization $\theta_0^{(i)}$ leads to a variety of "winning tickets" [103] whose parameters $\theta^{(i)}$ captures well the main modes of $p(\theta|\mathcal{D})$ (see discussion in section 2.5.2). The original "lottery ticket" hypothesis formulated by Frankle and Carbin in 2018 stipulates:

Lottery Ticket Hypothesis (Frankle and Carbin, 2018): *a randomly-initialized, dense neural network contains a sub-network that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

If we admit this hypothesis, then it becomes clear that different sub-networks emerge through different random initialization, each of them capturing various modes of $p(y|x, \mathcal{D})$ through $p(y|x, \theta^{(i)})$ (see Fig. 2.13). From this perspective, the next question is whether these winning tickets could be learnt directly from a single initialization θ_0, i.e. a single trained network f_θ could capture well all the modes in $p(y|x, \mathcal{D})$. Two solutions can be imagined: i) add a suitable regularization when minimizing the negative log-likehood of $p(y|x, \theta)$ (avoiding a small

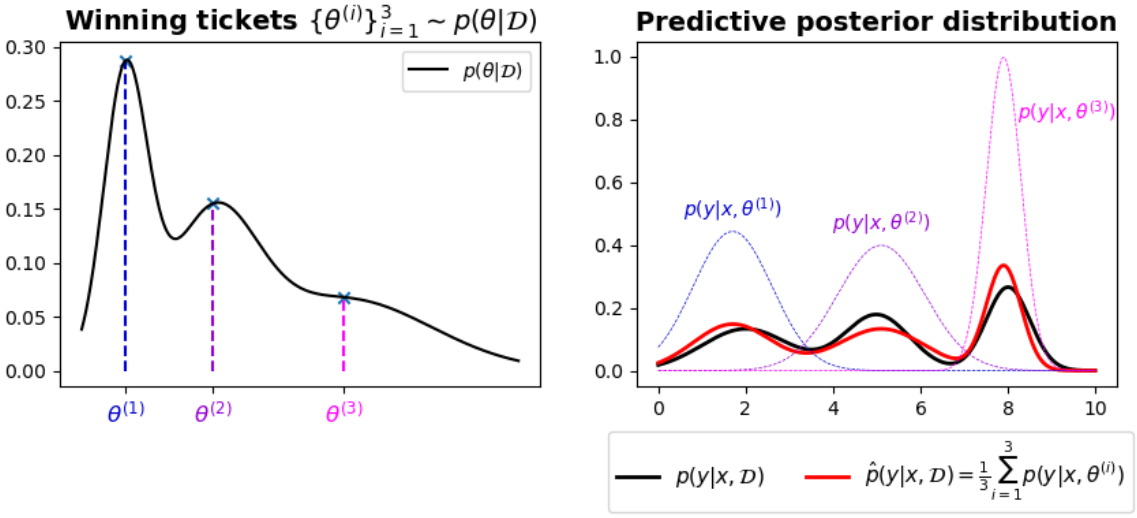


Figure 2.13: Deep ensemble learning captures different "winning tickets" $(\theta^{(i)})_{i\in[1..3]}$ (left pannel) leading to various distributions $p(y|x,\theta^{(i)})$ (right pannel) that approximate several modes of $p(y|x,\mathcal{D})$ (see eq. 2.8). Averaging these distributions allows to integrate both epistemic and aleatoric uncertainties, improving performance and calibration.

sub-network to win the lottery ticket too early and capture only partial modes of $p(y|x,\mathcal{D})$); ii) choose the initialization point θ_0 "carefully" (i.e. such that local minima close to θ_0 in the SGD optimization landscape provide a rich distribution $p(y|x,\theta)$ approximating well $p(y|x,\mathcal{D})$). The second solution is often referred to as Transfer Learning [44] and it will be discussed in the next chapter. Both solutions imply that the family of functions $\mathcal{F} = \{f_\theta\}$ is rich and expressive enough to have the existence of a single set of parameters θ such that: $p(y|x,\theta) \approx p(y|x,\mathcal{D})$.

Regarding MC-Dropout, similar conclusions stand (i.e.improved calibration and performance w.r.t baseline as the number of samples T increases), but the overall performance is somewhat lower than Deep Ensemble (in particular for sex prediction). It suggests that the different sub-networks obtained by activating dropout at test-time are not as diverse as the variety of winning tickets obtained from independently trained full networks.

The fundamental difference between MC-Dropout and Deep Ensemble is the variational distribution $q(\theta)$ chosen to approximate $p(\theta|\mathcal{D})$. In MC-Dropout, the variational distribution $q(\theta)$ induces strong correlations between the weights inside the network during training (i.e. strong redundancies between several sub-networks). On the other hand, in Deep Ensemble, these correlations are only induced by the training data $\mathcal{D}$: completely independent sub-networks may emerge for an over-parametrized network and a task with separate discriminative patterns. For instance, one sub-network may be specialized to extract gray matter atrophies in the temporal lobe while another network might discriminate hypertrophies in the prefontal lobe, both predictive of schizophrenia (and potentially depending on patient's age). Given such difference, we hypothesize that the difference in performance between MC-Dropout and Deep Ensemble is even more pronounced for smaller network. Indeed, in that regime, inducing strong redundancies between sub-networks could imply far less representative power for each one of

them.

Experiments on tiny-DenseNet

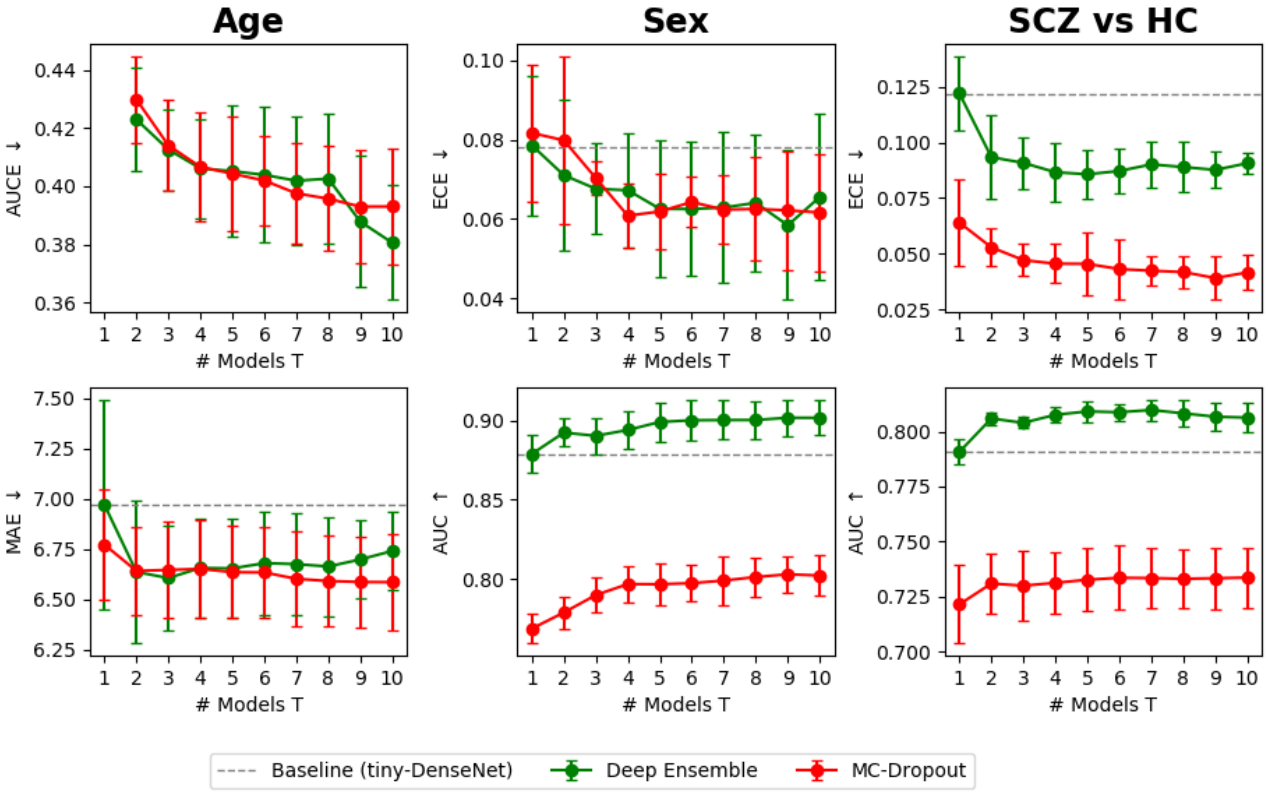


Figure 2.14: MC-Dropout clearly under-performs when reducing the model's capacity on all classification tasks compared to 1) baseline with a deterministic DNN and 2) Deep Ensemble integrating both aleatoric and epistemic uncertainty. However, when reducing model size, all models becomes well-calibrated on all tasks (especially compared to DenseNet121, see Fig. 2.12).

Results Fig. 2.14 confirms our hypothesis at least for classification tasks (SCZ vs HC and sex prediction). MC-Dropout clearly under-performs compared to Deep Ensemble while remaining well calibrated (even better than Deep Ensemble for SCZ vs HC). It suggests that the network makes more mistake than Deep Ensemble and it has little confidence in its prediction (as it should be). Overall, imposing strong redundancies in smaller network's weights hurts the performance. It supports our previous claim that MC-Dropout needs highly over-parametrized network to work well, suggesting that the variational distribution $q(\theta)$ is not adapted in that scenario. It is not the case for Deep Ensemble where it outperforms the baseline in all cases (in line with recent findings on semantic segmentation and depth completion [123]).

Interestingly, it is important to notice that all models (baseline, Deep Ensemble, MC-Dropout) are largely better calibrated with tiny-DenseNet than DenseNet (e.g. 12.5% vs 17.5% ECE for SCZ vs HC with baseline model, 4% vs 10% with MC-Dropout, 8% vs 17% ECE for sex classification with baseline, etc.). We retrieve the results obtained by Guo et al. in 2017[122] on standard vision datasets (e.g. CIFAR100 [179]) where modern highly over-parametrized

networks are too over-confident in their predictions. While we notice better calibration when applying MC-Dropout or Deep Ensemble techniques, there is still room for improvement.

Overall, this study highlights the importance of integrating deep uncertainty (aleatoric and epistemic) in DNN, especially when building computer-aided diagnosis tools. Not only it would allows clinicians to trust the AI system by giving a notion of confidence in the predictions made, but it also improve the overall performance of the algorithm itself. While Deep Ensemble gave the best trade-off between calibration and performance throughout our experiments, the current Bayesian theory opens new avenues for a wide range of variational distributions $q(\theta)$ (e.g. Gaussian dropout [174] instead of Bernouilli, etc.) that may be more suited for brain imaging data. Nonetheless, we acknowledge that the (implicit) distribution $q(\theta)$ underlying Deep Ensemble is still poorly understood theoretically and we hope that future research on the Lottery Ticket Hypothesis will enhance our comprehension.

Integrating deep uncertainty to outperform linear models

To conclude this section about deep uncertainty in DNN for brain imaging applications, we wanted to come back to the original question of this chapter: can we extract non-linear patterns inside brain imaging with DNN to outperform (regularized) linear models ?

To fairly answer to this question, we have integrated the main techniques presented in this chapter that led to an improvement in performance both for linear and non-linear models:

1. we use fully pre-processed VBM data for all models (instead or raw images, see section 2.3.2)
2. we use linear adjusted regression to remove site-related noise for linear regression and Kernel-SVM (see section 2.4.2)
3. we do not apply data augmentation (see section 2.4.1)
4. we model epistemic uncertainty inside DNN with Deep Ensemble
5. we use DenseNet121 backbone as encoder (see Fig. 2.2)

To limit the computational cost in these experiments, we limit the number of ensemble models to $T = 3$ (considering the results obtained in Fig. 2.12). We test all models on the 3 clinical tasks of interest (schizophrenia, bipolar disorder and ASD), again using the maximum number of available samples in BHB-10K ($N_{train} > 800$ for all tasks). All DNN are trained with a simple binary cross-entropy loss. and resuts are reported Table 2.8.

From Table 2.8, we observe that DNN is able to outperform linear models for 2 out of 3 classification tasks (bipolar disorder and ASD, the hardest tasks according to the average AUC between all models). Deep Ensemble provides large improvement for all tasks, confirming our previous analysis and the importance of integrating epistemic uncertainty inside DNN. Interestingly, for the "easiest" task among the three (schizophrenia detection), DNN are not

Task	Test Set	Deep Models		SML		
		Baseline	Deep Ensemble	rbf-SVM	Logistic ℓ_2	ElasticNet
SCZ vs HC ↑	Internal Test	$85.27_{\pm 1.60}$	$85.73_{\pm 0.53}(+0.46)$	$83.55_{\pm 0.00}$	$85.31_{\pm 0.07}$	$\mathbf{88.81_{\pm 1.03}}$
$N_{train} = 933$	External Test	$75.52_{\pm 0.12}$	$77.47_{\pm 0.71}(+1.95)$	$76.39_{\pm 0.00}$	$76.45_{\pm 0.15}$	$\mathbf{78.98_{\pm 0.98}}$
BD vs HC ↑	Internal Test	$76.49_{\pm 2.16}$	$\mathbf{79.49_{\pm 1.36}}(+3.00)$	$75.00_{\pm 0.00}$	$74.07_{\pm 0.09}$	$71.19_{\pm 2.29}$
$N_{train} = 832$	External Test	$68.57_{\pm 4.72}$	$\mathbf{76.11_{\pm 0.53}}(+7.54)$	$67.74_{\pm 0.00}$	$69.54_{\pm 0.33}$	$70.33_{\pm 2.47}$
ASD vs HC ↑	Internal Test	$65.74_{\pm 1.47}$	$\mathbf{67.67_{\pm 0.74}}(+1.93)$	$66.78_{\pm 0.00}$	$64.71_{\pm 0.22}$	$63.30_{\pm 4.78}$
$N_{train} = 1526$	External Test	$62.93_{\pm 2.40}$	$\mathbf{64.48_{\pm 1.51}}(+1.55)$	$59.10_{\pm 0.00}$	$63.98_{\pm 0.15}$	$57.98_{\pm 4.71}$

Table 2.8: Integrating epistemic uncertainty inside DNN through Deep Ensemble allows to outperform regularized linear models by a high margin on 2 out of 3 tasks (+6%/5% AUC for BD vs HC/ASD vs HC on external test). We reported average AUC for all models and the standard deviation by repeating each experiment three times. Baseline for DNN corresponds to a single DenseNet121 trained from scratch on VBM images. For Deep Ensemble, we aggregated three networks trained from different random initialization. Green numbers indicate improvement over DL baselines.

able to outperform ElasticNet (leading to very sparse solutions compared to logistic and rbf-SVM).

We attribute this to the "simplicity bias" [253] that occurs during training: DNN have an ability to over-fit rapidly on the "simplest features" (that can be less discriminative than more complex ones), leading to non-robustness in the solution found. Several evidence suggest such behavior: even after applying Deep Ensemble learning, a high performance gap between internal and external test is observed for schizophrenia detection, much more than the other 2 tasks (-8% AUC vs -3% AUC). Additionally, Deep Ensemble has a limited effect on performance, suggesting less diversity in the final representations, thus leading to poor generalization.

Perspectives. Quite recently, Teney et al. [274] proposed a new regularization term to evade the simplicity bias by training multiple MLP heads over a single encoder's backbone. All these heads are trained jointly with orthogonal gradient constraint during gradient-descent. This kind of methods can be formalized through Bayesian theory in the same way as Deep Ensemble (see previous section) and it offers an appealing training scheme to diversify the final DNN representations (notably here for schizophrenia detection). We have left this for future work.

2.6 Conclusion

In this chapter, we have investigated key properties of supervised DL models on anatomical brain imaging data. To conduct our analysis, we first have gathered a large collection of brain images through various sharing initiatives, leading to a large multi-site dataset. It notably includes patients with schizophrenia, bipolar disorder and autism but also a large cohort of healthy controls spanned from childhood to elder-hood.

From this dataset, we have shown that current SOTA DL models perform on par with regularized linear models at current clinical size for mental disorder classification tasks. They

tend to rapidly over-fit on noisy features (including site-related information), which notably prevents them from extracting additional geometrical discriminative patterns (e.g cortical foldings) buried inside raw images. We have observed such behavior repeatedly by analyzing their performance on external cross-site test sets and it shades light on an important bias in current neuroimaging datasets that will surely be amplified as more consortium initiatives arise. Interestingly, we have also demonstrated that DNN remain bias even as we reach the large-scale regime $N_{train} = 10$k for phenotype prediction, suggesting that "it's not all about larger dataset", as also illustrated on Alzheimer's disease by Varoquaux and Cheplygina [290].

From this analysis, we have studied data augmentation as regularization technique and data-based debiasing techniques (such as data harmonization) for DNN. We mostly found no improvement for the targeted clinical applications, suggesting that current augmentations crafted from the human perception need to be rethought for brain imaging.

Finally, as contemplated by Bzdok, Floris and Marquand [40], modelling biological variability and methodological uncertainty through Bayesian theory is urgently required for analyzing brain MRI in order to "*go beyond binary statements on existence vs non-existence of an effect and afford credibility estimates around all model parameters at play which thus enable single-subject predictions with rigorous uncertainty intervals.*" As a result, in the last section, we have used recent works on Bayesian DNN to model both aleatoric and epistemic uncertainties inside DNN, re-casting standard Dropout and Deep Ensemble techniques in this framework. We notably show significant improvement for both calibration and performance on all psychiatric disorder classification tasks with largely over-parameterized DNN. This work higlights the importance of modelling epistemic uncertainty and it opens up new avenues for developing new variational approximations of network's posterior distribution.

Chapter 3

Unsupervised representation learning for neuroimaging: a step towards transfer learning

Contents

This work has been presented in:

- **Contrastive Learning with Continuous Proxy Meta-Data for 3D MRI Classification**
  B. Dufumier, P. Gori, J. Victor, A. Grigis, E. Duchesnay et al.
  MICCAI, 2021
- **Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels**
  B. Dufumier, P. Gori, J. Victor, A. Grigis, E. Duchesnay
  NeurIPS Workshop on Medical Imaging Meets NeurIPS, 2021
- **Rethinking Positive Sampling for Contrastive Learning with Kernel**
  B. Dufumier, C. A. Barbano, R. Louiset, E. Duchesnay, P. Gori
  Submitted to ICML 2023

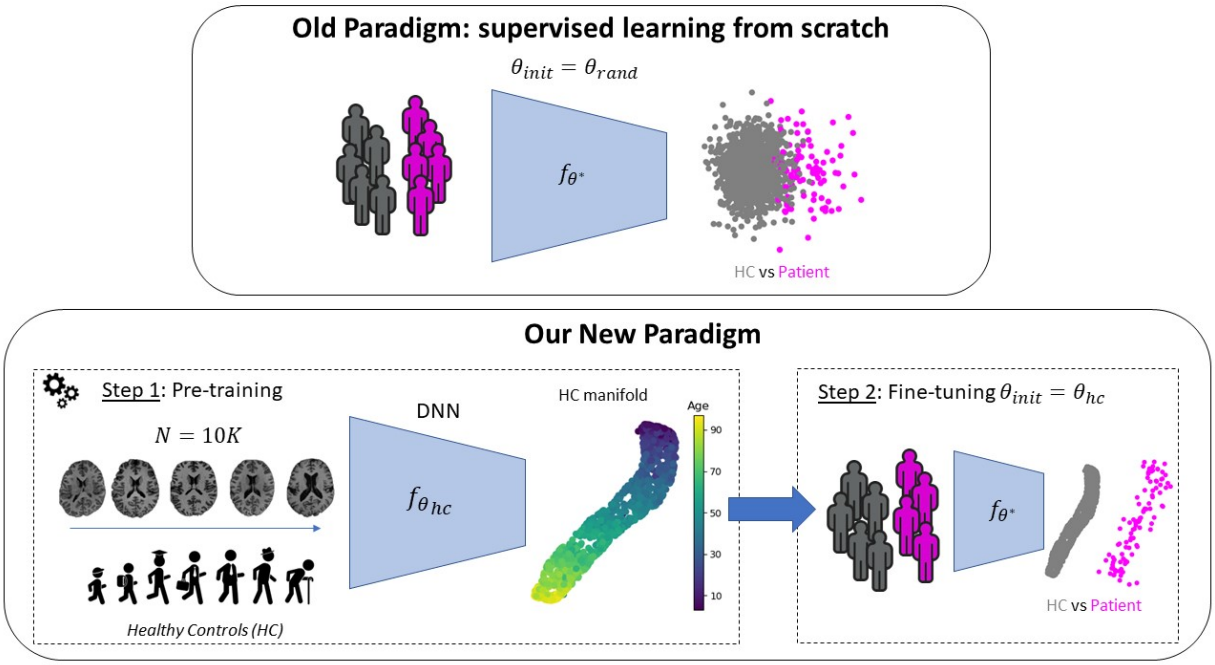


Figure 3.1: New paradigm for discriminating psychiatric disorders at the subject-level. In a pre-training phase, a non-linear DNN f_θ is trained to learn a low-dimensional embedding from a large brain imaging dataset of healthy controls, discovering the general variability associated with non-specific variables such as age and sex. This pre-training can be performed with self-supervised task (e.g. contrastive learning [52, 87]) or discriminative task (e.g. age prediction [25]). In a second step, the model is initialized with pre-trained weights $\theta_{init} = \theta_{hc}$ and fine-tuned to discriminate between patients and controls. Our main hypothesis is that the manifold learned during pre-training will allow easier discovery of the specific variability associated to the pathology of interest (e.g. abnormal cortical atrophy in temporal and pre-fontal regions for schizophrenia or ASD).

In the previous chapter, we aimed at discovering the representation capacity of DNN in a fully supervised context to discriminate between patients and controls. One of the main bottleneck of current DNN representation capacity is their need for (very) large dataset. It

was illustrated in the previous chapter on challenging classification tasks to detect psychiatric disorders where DNN struggled to find better solutions than linear models, given the current sample size ($N < 1$k).

Large population imaging initiatives such as the Human Connectome Project [286] (launched in 2010) or UKBioBank [37] (started in 2006, and imaging 100k subjects in the UK)-focused mainly on the healthy population - now enable the development of new AI tools for modelling the normal human brain development through life (from childhood to elder-hood). Discovering the data manifold from the healthy population allows notably to accurately model the biological variability inherent to healthy brains (e.g. related to phenotype/genotype information such as age, sex or genetics). From this perspective, pathological brains (e.g. from subjects with schizophrenia or bipolar disorder, who display abnormal cortical brain patterns compared to healthy groups) can be viewed as a deviation orthogonal to the tangent vector space of its unobserved "healthy twin", lying on the data manifold (as well illustrated by A. Aglinskas et al. in a recent Science article [4] focused on autism).

In this Chapter, we study how to model such low-dimensional manifold of the healthy population using self-supervised models based on contrastive learning. These discriminative models have several advantages over their generative counterparts (such as VAE [174] or GAN [116]): they do not need a computationally demanding pixel-level generation (which could be unnecessary for learning representations), they are easy to train and they do not explicitly model the data generating process but rather an approximation of its inverse [323]. We validate the models developed in this thesis on several clinical cohorts including patients with schizophrenia, bipolar disorder, autism but also Alzheimer's patients, thus covering a large spectrum of psychiatric and neurodenerative disorders.

In the first part, we present the original formulation of contrastive learning (CL) for visual representations [52, 124, 211] from an information theory point-of-view and we present its two main implementations with MoCo [136] and SimCLR [52]. As our first original contribution, we describe how auxiliary phenotyping information such as subject's age can be leveraged to shape the embedding space during optimization. This framework notably extends supervised contrastive learning to the weakly-supervised case using a similarity function between auxiliary signals. We also study several critical components of CL such as data augmentation and batch size and their impact on the final embedded representation.

In the second part, we provide an in-depth theoretical framework for CL. Based on this analysis, we ask whether data augmentation component (a critical component in today's CL models) can be partially removed in CL for learning visual representations in medical imaging. Accounting for the difficulty to find the relevant augmentations for medical datasets, we wonder whether generative models can serve as a prior to learn relevant representations. We develop a strong theory based on conditional kernel embedding and we demonstrate the utility of our framework on several toy examples and real-world brain MRI and chest X-ray scans.

3.1 Introduction to unsupervised representation learning

In the last chapter, we studied several supervised problems where we wanted to estimate the conditional distribution $p(y|x, \mathcal{D})$, given a training dataset $\mathcal{D}$ of labelled examples. Generally, the decision boundary separating examples of different classes can be learnt directly by optimizing a cross-entropy objective function. Self-supervised models are somewhat more general as they aim to learn a representation $z \in \mathbb{R}^d$ of the data $x \in \mathbb{R}^p$ ($d \ll p$) that can be used to study several supervised downstream tasks. That is, from the representation z, multiple labels can be "easily" inferred with different levels of granularity-for instance using only a linear combination of latent factors $(z_i)_{i\in[1..d]}$. If the data distribution $p(x)$ represents images of animals then the representation z should contain color, eyes' and ears' shape, whether it has a tail, its size etc. From this representation, the learnt representation allows to answer several questions: is this animal a dog or cat ? Is it a Dobermann or a Poodle ? Is it a baby Poodle or an adult ? Bengio et al. [30] identified several key factors for learning a "good representation" of the data:

- Expressiveness: "a reasonably-sized learned representation can capture a huge number of possible input configurations". In particular, this implies having a large number of features that can be re-used for a wide number of tasks. While each latent factor z_i can be independent from one another (e.g color vs shape), each of them can represent many different concepts. The number of concepts N can then be much bigger than the number of latent factors d;

- Disentangled factors: several factors of variation z_i should be independent from on another, thus respecting the hypothesis made by the neuroscientist Barlow [23] that the goal of sensory processing is to recode highly redundant sensory messages into a reduced factorial code, with independent components;

- Invariance: an abstract representation z of the data should be invariant to a high number of raw input variations (e.g. rotation/translation for object detection or illumination for scene detection). This can notably emerges with a high abstraction of the raw data as we go deeper for deep neural network architectures. In that case, the representation is less sensitive to the variation of a single raw input value (e.g. a pixel for images) and it can approximate highly non-linear functions of raw input. A good illustration is Convolutional Neural Network [189] where small translation invariance is encoded directly by design with pooling operator.

Current representation learning models can be viewed from three different perspectives [30]: 1) probabilistic models where the joint distribution $p(x, z)$ is modeled; 2) parametric mapping between input x and latent factors z (e.g. auto-encoders or self-supervised models) ; 3) manifold learning where data are assumed to lie on an implicit manifold where some variations (change of illumination, pose, etc.) of an input x is traduced by variations along tangent vectors of this

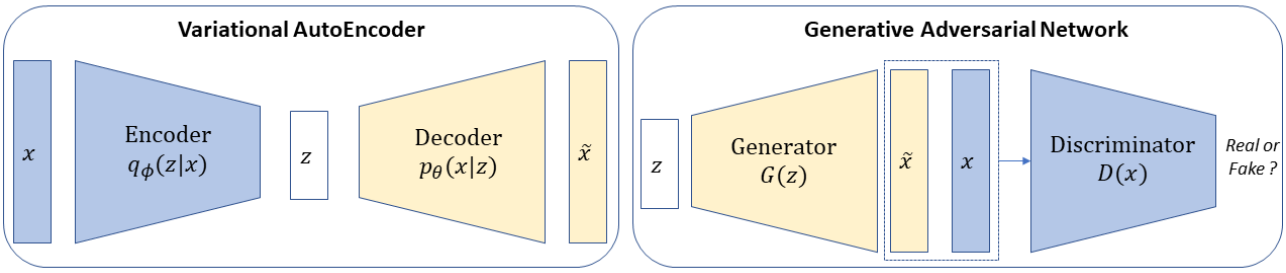


Figure 3.2: Overview of two main deep generative models for representation learning: VAE and GAN. Both models have numerous variants for specific applications but the main original ideas are depicted.

point in the manifold. Current state-of-the-art generative models (GAN [116] and VAE [173]) use a probabilistic model $p(x, z)$ to model the true data distribution $p(x)$ by setting a prior $p(z)$ and by approximating the conditional distribution $p(x|z)$. In fact, classical Principal Component Analysis (PCA) can also be seen as a simple probabilistic model where $p(x|z)$ is explicitly estimated with a Gaussian distribution [277]:

$$p(z) = \mathcal{N}(z; 0, \mathbf{I}) \tag{3.1}$$

$$p_\theta(x|z) = \mathcal{N}(x; Wz + \mu, \sigma^2\mathbf{I}) \tag{3.2}$$

where $\theta = \{W, \mu, \sigma^2\}$. The classical loading factors in PCA span the same space as the p columns of W (reminding that $x \in \mathbb{R}^p$), estimated by maximum likelihood[1] While PCA is generally the simplest (and the oldest) model for representation learning, it can be formalized from the three perspectives above-mentioned (probabilistic, parametric as linear autoencoder, manifold learning [30]). Thus, it gives a way to connect these three point-of-view in a simple manner.

We first start by giving an overview of generative models for representation learning, including GAN and VAE. Then, we continue by giving an in-depth analysis of current state-of-the-art self-supervised contrastive algorithms, based on instance discrimination and widely used for learning visual representations. In section 3.3, we shall present a connection between generative and self-supervised contrastive learning for learning representations.

3.1.1 A little journey with deep generative models

Deep generative models are a family of generative models that learn the true data distribution $p(x)$ with deep neural networks. They give insight about the true factor of variations underlying the data generative process by explicitly approximating the conditional distribution $p(x|z)$, given a prior distribution $p(z)$. They are also well-known for their numerous real-world applications including (but not limited to): super-resolution, style-transfer [166], image-to-image translation [155], class-conditional generation [201], image denoising, disentanglement [56], pre-

[1]However, these columns are not necessarily orthonormal.

training. We describe hereafter two main models used also for representation learning: VAE and GAN.

Variational AutoEncoder

VAE assumes that the data are coming from an underlying unobserved latent variable z, explaining the observed input x. Mathematically, it assumes that it exists a joint distribution $p(x, z)$ between a high-dimensional input $x \in \mathbb{R}^p$ and its low-dimensional representation $z \in \mathbb{R}^d$ ($d \ll p$). To keep tractable expressions, both distributions $p(z)$ (the prior) and $p_\theta(x|z)$ (parametric conditional distribution) are assumed to be Gaussian:

$$p(z) = \mathcal{N}(z; 0, \mathbf{I}) \tag{3.3}$$

$$p_\theta(x|z) = \mathcal{N}(x; f_\theta(z), \sigma^2\mathbf{I}) \tag{3.4}$$

In the above expression, we started to introduce a DNN called *decoder* f_θ mapping a latent vector z to some realistic input x. Please note the similarity between this model and probabilistic PCA mentionned above. However, the likelihood $p_\theta(x) = \int p_\theta(x|z)p(z)dz$ is often intractable or it requires a costly MCMC sampling[2]. Instead, VAE uses variational inference (VI) to approximate the "reversed" distribution $p_\theta(z|x)$.

VI introduces a tractable variational distribution $q_\phi(z|x)$ approximating $p_\theta(z|x)$ (again, intractable since it requires to compute $\int p_\theta(x|z)p(z)dz$). This distribution is learned by an *encoder* $e_\phi(x) = [\mu(x), \sigma^2(x)]$ that parametrizes $q_\phi(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x))$. In the end, VAE optimizes a lower bound of the likelihood $p_\theta(x)$, called Evidence Lower Bound (ELBO):

$$p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - KL(q_\phi(z|x)||p(z)) = -\mathcal{L}_{ELBO} \tag{3.5}$$

In practice, the latent representation z can be easily obtained from a VAE since the (approximated) distribution of $p_\theta(z|x)$ is available. Importantly, during training, encoder e_ϕ and decoder f_θ are trained jointly to minimize $\mathcal{L}_{ELBO}$.

Generative Adversarial Network

Original formulation. Originally, GAN is inspired by Noise Contrastive Estimation (NCE) [124] which aims at learning the true data distribution $p(x)$ with a parametric distribution $p_\theta(x)$. In NCE, the model learns to discriminate between true data examples and noise using a logistic function. In other terms, it learns a parametric model $p_\theta(x)$ by comparing a set of training examples (sampled from $p(x)$) with another set of noise examples (sampled from $p_{noise}(x)$). It can be showed [124] that it leads to a consistent estimator θ^* such that $p_{\theta^*}(x) = p(x)$ under mild assumptions. However, in practice, one issue arises when the model rapidly distinguishes true examples from noise, using only a very rough approximation of the true data distribution $p(x)$ and a few training examples.

[2]PCA was simpler in that regard since f_θ was a linear mapping and all mathematical expressions were tractable.

In GAN, the idea is quite similar but rather than using a fixed noise distribution $p_{noise}(x)$, it is learned through a DNN. More precisely, a generator $G(z)$ learns to generate realistic samples from a Gaussian prior $p(z) = \mathcal{N}(z; 0, \mathbf{I})$, while a discriminator $D(x)$ learns to distinguish between true samples $x \sim p(x)$ and "fake" generated ones $x \sim p_g(x)$ where p_g is the fake distribution induced by G. This way, the optimization problem is a min-max objective:

$$\min_G \max_D \mathbb{E}_{p(x)} \log D(x) + \mathbb{E}_{p(z)} \log(1 - D(G(z))) \tag{3.6}$$

At optimum, it can been shown that the optimal discriminator is reached when $D^*(x) = \frac{p(x)}{p(x)+p_g(x)}$ (for a fixed G) and $p_g(x)$ converges to $p(x)$ for large enough capacity discriminator and generator.

This formulation is fairly general and does not specify the exact architecture of generator and discriminator. A lot of works have extended this original idea by providing: a suitable architecture for D and G (e.g. DCGAN [226] for convolutional models), an improved objective function for stability (e.g. Wasserstein-GAN GP [121]), an encoder $E(x)$ to learn the reverse mapping between an input x and its latent representation z, thus modelling $p(z|x)$ (e.g. ALI [88], BiGAN [85] and BigBiGAN [84]). All these improvement led to state-of-the-art generative models capable of generating high-quality (and fidelity) images as well as high-quality representations (see hereafter). Nonetheless, they still under-perform compared to self-supervised models for representation learning and their training require a massive amount of hyper-parameter tuning (and engineering tricks).

BiGAN and ALI for representation learning. As the reader may have noticed, the original formulation only estimates $p(x|z)$ for a prior $p(z) = \mathcal{N}(z; 0, \mathbf{I})$ and then implicitly learns $p(x)$ through $p_g(x) = \mathbb{E}_{p(z)} p_\theta(x|z)$. It does not learn the "reverse" distribution $p(z|x)$, mapping back an input x to a latent vector z (which can be used for representation learning). As a result, it avoids the need of a variational distribution $q_\phi(z|x)$ estimating $p(z|x)$ as in VAE.

In BiGAN and ALI, both $p(x|z)$ *and* $p(z|x)$ are estimated using two deep networks (as in VAE): an encoder $q_\phi(z|x)$ (mapping x to z with a generator $G_z(x)$) and a decoder $p_\theta(x|z)$ (mapping z to x with a generator $G_x(z)$). They induce two joint distributions $q_\phi(x, z) = q_\phi(z|x)p(x)$ and $p_\theta(x, z) = p_\theta(x|z)p(z)$ where $p(z) = \mathcal{N}(z; 0, \mathbf{I})$ and $p(x)$ is the true data distribution. Using the exact same idea as original GAN, a discriminator $D(x, z)$ is trained to distinguish between a "true" pair $(x, \hat{z}) \sim q_\phi(x, z)$ and a fake pair $(\hat{x}, z) \sim p_\theta(x, z)$. As one can expect, after training, distributions $p_\theta(x, z)$ and $q_\phi(x, z)$ are supposed to match in order to fool the discriminator. The objective function is almost identical to GAN:

$$\min_G \max_D \mathbb{E}_{(x,\hat{z})\sim p_\theta(x,z)} \log D(x, \hat{z}) + \mathbb{E}_{(\hat{x},z)\sim q_\phi(x,z)} \log(1 - D(\hat{x}, z)) \tag{3.7}$$

Following the same theoretical work as in GAN, it can be shown that, for a fixed generator, the optimal discriminator is $D^*(x, z) = \frac{p_\theta(x,z)}{p_\theta(x,z)+q_\phi(x,z)}$. More interestingly, for an optimal discrim-

inator $D^*(x, z)$, the optimal generator reaches it minimum if, and only if, $p_{\theta^*}(x, z) = q_{\phi^*}(x, z)$. In particular, it means that the two marginal distributions are equal so $q_{\phi^*}(z) = \mathcal{N}(z; 0, \mathbf{I})$ and $p_{\theta^*}(x) = p(x)$ at optima. Finally, at optima, the two conditional generators $G_z(x)$ and $G_x(z)$ are the inverse of each other under mild assumption: $G_z^{-1} = G_x$ and $G_x^{-1} = G_z$.

In 2019, Donahue and Simonyan [84] achieved state-of-the-art results for both image generation and representation learning on ImageNet using this model (with some modifications to generator and more regularization terms in the loss). It notably suggests that a representation learning objective improves the generative process. Interestingly, a more recent study [51] also suggested the opposite: better generative models learn better representations at least for visual representations using GPT-2 model. Note that we refer to "better representation" according to the linear probing tool: the representation quality is measured by the predictive power (e.g. accuracy) of a logistic regression trained to predict labels on a given downstream task from the model's representation.

Conclusion. All the current generative models used for representation learning approximate an unknown distribution $p(x|z)$, conditional distribution of true observed data x from latent (unobserved) variable z. Nevertheless, the primary purpose of representation learning is to learn the reverse mapping: from an observed input x, we would like to encode its compressed latent code z that respects the 3 main principles explained above. Now, the question that self-supervised contrastive models try to answer is: can we avoid estimating $p(x|z)$ for learning representation $p(z|x)$ with a discriminative models ?

3.1.2 Self-supervised contrastive learning

Self-supervised models fall into the category of discriminative models that does not rely on labeled examples to learn data representation, as it is classically done for supervised learning. They do not learn a mapping between an input x (either images, text, speech, etc. or a combination) and some pre-defined human annotation y–which has a finite level of granularity and only depicts some features about x–but rather they intend to learn a generalizable representation z of x that extracts high-level abstract features that we, as humans, also use to describe x. They are built through the main principles formalized by Bengio et al. in its survey on representation learning [30], in particular following *expressiveness* (or distributed) and *invariance* principles as we shall see. We will focus on instance-based discriminative models since they represent the large majority of state-of-the-art models for representation learning of visual representations.

Contrastive learning from an information bottleneck perspective

Supervised setting. In supervised learning, one wants to learn relevant information about input random variable $X \in \mathcal{X}$ giving as much information as possible about its label $Y \in \mathcal{Y}$ (also seen as a random variable). The relationship between X and Y is modelled as a joint distribution $p(X, Y)$ and the amount of information Y gives about X is the Mutual Information

(MI) $I(X,Y) = KL(p(X,Y)||p(X)p(Y))$. A classical view of DNN is that it learns a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$ in a layer-wise manner, in particular mapping X to an intermediate representation Z_θ, mapped to its final prediction $Y_\theta = f_\theta(X)$. Since $I(X,Y) \geq I(g(X),Y)$ for any function g (i.e. $g(X)$ cannot contains more information about X than X itself) then:

$$I(X,Y) \geq I(Z_\theta,Y) \geq I(Y_\theta,Y) \tag{3.8}$$

Equality in previous inequalities is equivalent to Z_θ and Y_θ are sufficient statistics of X for Y (also equivalent to $I(X,Y|Z_\theta) = I(X,Y|Y_\theta) = 0$). Optimization of DNN thus implies a compression phase-where each layer removes irrelevant information about X–and a predictive phase, where Y_θ retains as much information as possible about Y, formally trying to minimize $I(X,Z_\theta)$ while keeping $I(Z_\theta,Y)$ maximized.

Unsupervised learning. Now, we assume that we do not have access to ground-truth label Y during training anymore. Nevertheless, we still want to learn a representation Z_θ that is as much informative about Y as possible. Please note that Y is not necessarily human annotations anymore but it can rather be some fine-grained semantic properties about input X (such as brain shape in temporal or prefontal lobe for brain imaging data). The key idea behind contrastive learning is again based on NCE (see below): to learn whether given samples are sampled from a distribution p^+ (a.k.a *positive* distribution) or from another distribution p^- (a.k.a *negative* distribution), thus learning implicitly p^+.

In particular, let us consider two random variables V_1 and V_2, representing two random views of the same input X, e.g. two different parts of the same image (see below for a formal definition). We will make two strong assumptions about V_1 and V_2, which will be discussed after:

- (Label preserving) V_1 and V_2 are sufficient statistics of X for Y, i.e. $I(V_1,Y) = I(V_2,Y) = I(X,Y)$
- (Strict Redundancy) V_1 and V_2 share only label information Y: $I(V_1,V_2) = I(V_1,Y) = I(V_2,Y)$

From these two assumptions, it is easy to see that $I(X,Y) = I(V_1,V_2)$ so now the problem consists in preserving all shared information between V_1 and V_2 from the representations $Z^1_\theta = f_\theta(V_1)$ and $Z^2_\theta = f_\theta(V_2)$. One way to do it is by training a *critic* $E_\theta(V_1,V_2) = -\frac{Z^1_\theta \cdot Z^2_\theta}{||Z^1_\theta|| \cdot ||Z^2_\theta||}$ (viewed as an energy function) such that it gives low values to plausible pairs $(v_1,v_2) \sim p(V_1,V_2)$ (=positive distribution) and high values to implausible pairs $(v_1,v_2) \sim p(V_1)p(V_2)$ (=negative distribution). The following InfoNCE [211] MI estimator can be used to train such critic:

$$I_{NCE}(V_1,V_2) = \mathbb{E}_{(v_1^i,v_2^i)_{i\in[1..N]} \sim p(V_1,V_2)} \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{-E_\theta(v_1^i,v_2^i)}}{\frac{1}{N}\sum_{k=1}^{N} e^{-E_\theta(v_1^i,v_2^k)}} \right) \tag{3.9}$$

Here N designates the number of pairs (v_1^i, v_2^i) used to estimate $I(V_1, V_2)$. It is worth noting two interesting properties about this MI estimator [222]:

1. (Consistency) InfoNCE converges to true MI: $I_{NCE}(V_1, V_2) \xrightarrow[N\to\infty]{} I(V_1, V_2)$ for an optimal critic $E_{\theta^*}(V_1, V_2) = -\log p(V_2|V_1) - \alpha(V_2)$ where $\alpha(\cdot)$ is an arbitrary function;

2. (Boundness) InfoNCE is upper bounded by $\log(N)$ and the true MI: $I_{NCE}(V_1, V_2) \leq \min(I(V_1, V_2), \log(N))$

Point 2 notably means that, in a real-world scenario, the number of pairs N to draw may need to be large if the MI to estimate is large. It also justifies why we can seek to maximize such estimator to find an optimal critic E_{θ^*} using InfoNCE loss:

$$\mathcal{L}_{InfoNCE} = -I_{NCE}(V_1, V_2) \tag{3.10}$$

Connection to NCE. As the reader may have guessed, InfoNCE is inspired from NCE formulation. To establish the connection, let's consider N samples $(v_1^i, v_2^i)_{i\in[1..N]} \overset{\text{iid}}{\sim} p(V_1, V_2)$. From NCE perspective, each (v_1^i, v_2^i) is considered as "observed data" and (v_1^i, v_2^j) for $j \neq i$ as the reference "noise" data (following $p(V_1)p(V_2)$). The energy function $E_\theta(v_1, v_2)$ implicitly defines a parametric distribution $p_\theta(v_1, v_2)$ such that:

$$E_\theta(v_1, v_2) = -\log p_\theta(v_1, v_2) + \log p(v_1)p(v_2) \tag{3.11}$$

To accurately estimate $p(v_1, v_2)$ using $p_\theta(v_1, v_2)$, NCE uses a logistic regression $h_\theta(v_1, v_2) = \frac{1}{1+e^{E_\theta(v_1,v_2)}}$ to tell whether each pair (v_1^i, v_2^j) is either sampled from $p(V_1, V_2)$ (positive distribution) or $p(V_1)p(V_2)$ (negative distribution). The final NCE loss is a simple (weighted) binary cross-entropy loss [31]:

$$\mathcal{L}_{NCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log h_\theta(v_1^i, v_2^i) + \frac{1}{N-1}\sum_{j\neq i}\log(1 - h_\theta(v_1^i, v_2^j))\right) \tag{3.12}$$

NCE guarantees that optimizing $\mathcal{L}_{NCE}$ w.r.t θ leads to a consistent estimator $p_{\theta^*}(V_1, V_2)$ of $p(V_1, V_2)$ (under mild assumptions, see [124]). We can prove (see Appendix B.1) that InfoNCE loss upper bounds NCE:

$$\mathcal{L}_{NCE} \leq \mathcal{L}_{InfoNCE} + \log(1 + e) + O\left(\frac{1}{N}\right) \tag{3.13}$$

Thus minimizing InfoNCE loss should also minimize NCE loss to some extent and it draws a connection between the original NCE formulation and the current InfoNCE implementation used in practice.

All this theory is based on the two assumptions about views V_1 and V_2 (namely label preservation and strict redundancy), that connect the (unobserved) semantic label Y with the the input

data X we have. It is appealing for its simplicity but it requires the actual practical definition of V_1 and V_2. All the following practical applications that use this theory perform "**instance-based discrimination**" [307], meaning they try to recover "real" pair $(v_1, v_2) \sim p(V_1, V_2)$ representing various aspects of the same underlying instance x (image, text, audio, etc.), from "fake" ones $(v_1, v_2) \sim p(V_1)p(V_2)$.

Invariance principle. The key idea behind contrastive learning based on instance discrimination is invariance, one of the three main principles identified by Y. Bengio [30]. Intuitively, we are trying to learn the semantic content Y of an input X (invariant across views), independent of its style S representing irrelevant change in the input (as formally defined in [292]). Both Y and S are latent factors and causally produce the observed input X but we are only trying to discover the (relevant unobserved) content Y through the MI tool $I(V_1, V_2)$.

Instance discrimination models

The previous model has been popularized with two successful implementations, SimCLR [52] and MoCo [136] for learning visual representations in 2020. Both use the InfoNCE loss (or a closed form) to learn the representations on ImageNet [74]. They introduce all the main components to perform unsupervised contrastive learning as we currently know and they have been used for various derived applications: semi-supervised learning [53], transfer learning [52], etc.

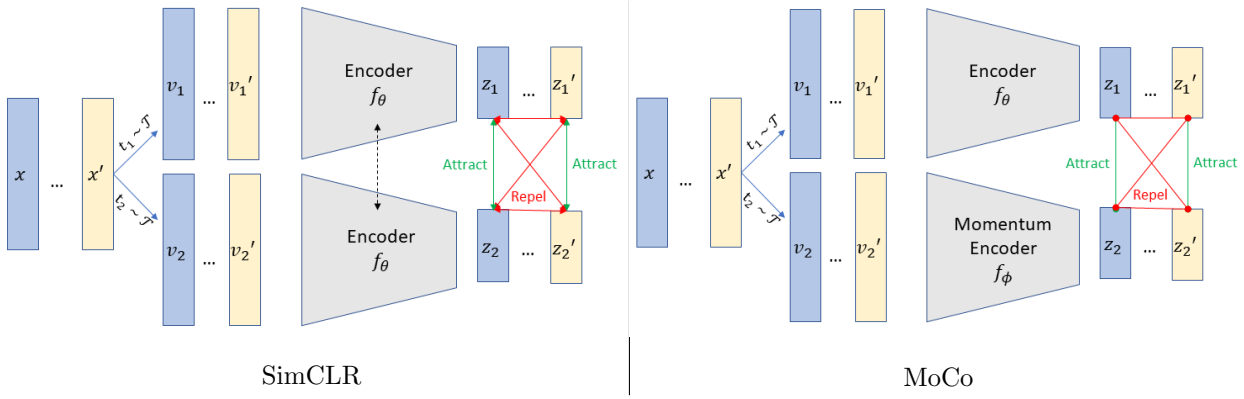


Figure 3.3: Two implementations (SimCLR and MoCo) of contrastive learning for unsupervised representation learning based on instance discrimination. While SimCLR uses the same encoder f_θ to map views v_i, MoCo uses an momentum encoder f_ϕ during training. Both encode two views (v_1, v_2) of the same instance x, obtained by sampling according to a set of transformations $t_i \sim \mathcal{T}$, $v_i = t_i(x)$. InfoNCE loss is used to attract (v_1, v_2) while repelling uniformly all views from all instances in a batch.

SimCLR [52] This method essentially introduces a simple and scalable learning algorithm by defining views V_1 and V_2 from a strong data augmentation strategy involving 10 different transformations (see Fig. 3.3), the two most important being random crop and color jittering. Formally, it introduces a set of transformations $\mathcal{T}$ which induces two random variables $V_1 =$

$T_1(X)$ and $V_2 = T_2(X)$ with $T_1, T_2 \sim \mathcal{T}$[3]. The InfoNCE loss is used during training along with a single encoder $f_\theta : \mathcal{X} \mapsto \mathcal{Z}$.

To comply with the label preserving and strict redundancy hypothesis made in the previous section, the random transformations T_1 and T_2 applied to X must: i) not be too strong in order to preserve the semantic label Y; ii) neither too light to share only the semantic content inside X. This has been well illustrated by Tian et al. [276] where he empirically showed a "sweet spot" for transformations T_1 and T_2 that are neither too strong or too light in order to comply: $I(V_1, V_2) \approx I(X, Y)$.

Finally, SimCLR [52] also demonstrates that InfoNCE loss needed a large batch size ($N > 8$k) to achieve the best results along with a non-linear "projection head" (corresponding in practice to a small MLP added on top of the encoder f_θ during training) that is thrown at test-time. While the former result was somehow expected by previous theory (since the MI estimator $I_{NCE}(V_1, V_2)$ is bounded by $\log(N)$), the latter is more surprising and it still does not have satisfactory explanations. One interesting observation made in [52] is the level of invariance to rotation and Sobel filtering captured before and after the projection head: the final representation (after projection) is more invariant to these transformations than before, which suggest that augmentations applied may be too strong (thus not preserving completely semantic content Y) but inductive bias through network's architecture may prevent representation collapse.

MoCo [136] (v1, v2, v3). The original idea is to maintain a large queue of latent representations z during training in order to increase drastically the number of "negative pairs" (i.e. fake pairs (v_1^i, v_2^j) for $j \neq i$ sampled from the negative distribution $p(V_1)p(V_2)$) during training, while optimizing InfoNCE loss. As a result, it avoids the computational burden of SimCLR since it does not require a large batch size, while it still estimate $I(V_1, V_2)$ with a large N. To do so, it relies on the previous representations computed during the last iterations to maintain and update the queue. A momentum mechanism is also introduced, originally because it did not use data augmentation to create views so they needed another mechanism to perform contrastive learning (data augmentation was introduced in MoCov2). It basically consists in introducing another encoder f_{θ_2} initially independent from f_{θ_1} but whose weights are updated slowly according to: $\theta_2 \leftarrow m\theta_1 + (1-m)\theta_2$ with $m \in [0, 1[$ a hyper-parameter close to 1. While no "views" are actually introduce in original MoCo, it is relying entirely on DNN architecture to produce pair of representations $(f_{\theta_1}(v_1), f_{\theta_2}(v_2))$, defining the energy function $E_\theta(v_1, v_2) = f_{\theta_1}(v_1) \cdot f_{\theta_2}(v_2)$ where $\theta = \{\theta_1, \theta_2\}$. It is worth noting that in the last version (MoCov3 [57]), the original queue used in MoCo was removed, making use only of large batch size, strong augmentations and the momentum mechanism with Vision Transformer backbone.

[3] We use an abuse of notations since T_1 and T_2 are random variables that indicate the augmentations to apply along with their strength and $T_i(X)$ hides the actual application of the selected augmentations $t_i \sim p(T_i)$ to X with a deterministic mapping $g : (\mathcal{X}, \mathcal{T}) \to \mathcal{X}$.

To contrast or not contrast ?

Contrastive learning is fundamentally based on NCE idea, that is learning to recognize if samples are drawn from a distribution p^+ (often referred to as *positive* distribution) or a distribution p^- (referred to as *negative* distribution). The previous theory was based on mutual information tool introduced by information theory (a point-of-view that dates back to 1992 [27]). Nevertheless, it is currently unclear whether this negative distribution p^- is required to learn representations. From an energy-based (EBM) point of view, the previous model imposes low energy values to positive pairs and high values to negative pairs. However, as Y. LeCun states [187]: "*when x is in a high-dimensional space, and if the EBM is flexible, it may require a very large number of contrastive samples to ensure that the energy is higher in all dimensions unoccupied by the local data distribution. Because of the curse of dimensionality, in the worst case, the number of contrastive samples may grow exponentially with the dimension of the representation. This is the main reason why I will argue against contrastive methods*".

Nonetheless, NCE taught us that it is by comparing two distributions that we can learn about one of them. If we assume to only know positive pairs $(v_1^i, v_2^i)_{i\in[1..N]} \sim p(V_1, V_2)$ then how can we learn this joint distribution ?

Following the previous EBM point-of-view, the simplest way to do it is by optimizing the negative log-likelihood of the energy model $p_\theta(V_1, V_2) = \frac{\exp(-E_\theta(V_1,V_2))}{Z_\theta}$ where $Z_\theta = \int \exp(-E_\theta(v_1, v_2))dv_1dv_2$ through a "Non-Contrastive" loss:

$$\mathcal{L}_{NC} = -\sum_{i=1}^{N} \log p_\theta(v_1^i, v_2^i) = \sum_{i=1}^{N} (E_\theta(v_1^i, v_2^i) + \log Z_\theta) \tag{3.14}$$

As for all EBM, the main difficulty is to estimate Z_θ (usually performed with Monte-Carlo Markov Chain using Langevin dynamics). However, two main methods have alleviate the need for estimating Z_θ by using other "tricks" that are still currently poorly understood[4]. Importantly, we know that optimizing only the energy function $E_\theta(v_1, v_2)$ using positive pairs $(v_1, v_2) \sim p(V_1, V_2)$ leads to a representation collapse (where all input x are mapped to the same representation z).

BYOL [120] and SimSiam [55]. Both these methods optimize implicitly an energy function but with 2 main tricks to avoid the representation collapse: 1) a projection head h_{θ_1} (e.g. small MLP) and 2) a "stop-grad" function during optimization. Additionally, BYOL uses a momentum mechanism that SimSiam removes, showing it is not the main component leading

[4] At the time of writing, there has been no intent to directly use MCMC sampling to optimize this "non-contrastive" objective, thus this view is still exploratory considering our current knowledge.

to state-of-the-art performance. In particular, SimSiam uses the following loss function[5]:

$$\mathcal{L}_{SimSiam} = -\sum_{i=1}^{N} \frac{h_{\theta_1}(f_{\theta_2}(v_1^i))}{||h_{\theta_1}(f_{\theta_2}(v_1^i))||} \cdot \text{stopgrad}\left(\frac{f_{\theta_2}(v_2^i)}{||f_{\theta_2}(v_2^i)||}\right) \tag{3.15}$$

where f_{θ_2} designates the encoder. This "stopgrad" function is fundamentally related to the way the loss is optimized (through Stochastic Gradient Descent). It means that the gradient is not back-propagated to update the weights θ_2 through the right-hand side of eq. 3.15. Understanding what is the underlying energy function implicitly optimized and why it does not collapse is still an open problem.

Barlow Twins and the redundancy principle [314] A different line of work explored self-supervised learning using redundancy reduction principle. It hypothesizes that each input can be represented by a compressed code with statistically independent components (following the second concept introduced by Y. Bengio [30]). This constraint avoids the need for a contrastive term and it is implemented in practice with a penalization on non-diagonal terms of the (empirical) cross-correlation matrix between $Z_\theta^1 = f_\theta(V_1)$ and $Z_\theta^2 = f_\theta(V_2)$. Estimation of this matrix can be performed by sampling only positive samples $(v_1^i, v_2^i) \sim p(V_1, V_2)$.

Clustering-based approach. Finally, SwAV [43] introduces a new clustering-based approach for learning representations. Instead of imposing close representations between views v_1 and v_2 of an instance, it enforces similar assignment between $f_\theta(v_1)$ and $f_\theta(v_2)$ to prototype vectors or *codes* (viewed as centroids in a clustering problem). This strategy has two fundamental differences with instance-based discrimination: i) it assumes the existence of a finite codebook (*i.e.,* set of prototypes) to which all representations can be map (either with hard or soft assignment); ii) it does not compare views representation directly. In instance-based discrimination, each image can have its own code, potentially independent from any other image codes, letting the possibility of an infinite codebook. SwAV instead assumes that all input representations may be described by a *finite* codebook (thus taking a step towards symbolic structure [260], yet without any notion of compositionality). Both SwAV and "SimCLR-like" approaches result in state-of-the-art results and it is not clear whether the use of a finite codebook is mandatory for representation learning.

3.2 Contrastive learning with auxiliary information

3.2.1 Context

Recently, self-supervised representation learning methods have shown great promises, surpassing traditional transfer learning from ImageNet to 3D medical images [320]. These models can

[5] We omitted the symmetrization term in this expression for clarity.

be trained without costly annotations and they offer a great initialization point for a wide set of downstream tasks, avoiding the domain gap between natural and medical images. They mainly rely on a pretext task that is informative about the prior we have on the data. This proxy task essentially consists in corrupting the data with non-linear transformations that preserve the semantic information about the images and learn the reverse mapping with a Convolutional Neural Network (CNN). Numerous tasks have been proposed both in the computer vision field (inpainting [216], localization of a patch [83], prediction of the angle of rotation [111], jigsaw [208], etc.) and also specifically designed for 3D medical images (context restoration [49], solving the rubik's cube [321], sub-volumes deformation [320]). They have been successfully applied to 3D MR images for both segmentation and classification [270, 272, 320, 321], outperforming the classical 2D approach with ImageNet pre-training. Concurrently, there has been a tremendous interest in contrastive learning [126] over the last years. Notably, this unsupervised approach almost matches the performance over fully-supervised vision tasks and it outperforms supervised pre-training [43, 52, 136].

Intuition. A single encoder is trained to map semantically similar "positive" samples close together in the latent space while pushing away dissimilar "negative" examples. In practice, all samples in a batch are transformed twice through random transformations $t \sim \mathcal{T}$ from a set of parametric transformations $\mathcal{T}$. For a given reference point (anchor) x, the positive samples are the ones derived from x while the other samples are considered as negatives. Most of the recent works focus in finding the best transformations $\mathcal{T}$ that degrade the initial image x while preserving the semantic information [52, 276] and very recent studies intend to improve the negative sampling [61, 234]. However, two different samples are not necessarily semantically different, as emphasized in [61, 302], and they may even belong to the same semantic class. Additionally, two samples are not always equally semantically different from a given anchor and so they should not be equally distant in the latent space from this anchor.

Contributions. In this work, we assume to have access to *auxiliary information* containing relevant information about the images at hand (*e.g* participant's age). We want to leverage these *auxiliary information* during contrastive learning in order to build a compressed representation of our data preserving these information. To do so, we propose a new y-Aware InfoNCE loss inspired from the Noise Contrastive Estimation loss [124] that aims at improving the positive sampling according to the similarity between two auxiliary information. Differently from [170], i) we perform contrastive learning with continuous auxiliary a (not only categorical) and ii) our first purpose is to train a generic encoder that can be easily transferred to various 3D MRI target datasets for classification or regression problems in the very small data regime ($N \leq 10^3$). It is also one of the first studies to apply contrastive learning to 3D anatomical brain images [47]. Our main contributions are:

- we propose a novel formulation for contrastive learning that leverages *continuous* auxil-

iary information and derive a new loss, namely the y-Aware InfoNCE loss, generalizing supervised contrastive loss;

- we empirically show that our unsupervised model pre-trained on a large-scale multi-site 3D brain MRI dataset comprising $N = 10^4$ healthy scans reaches or outperforms the performance of CNN model fully-supervised on 3 classification tasks under the linear protocol evaluation;
- we demonstrate that our approach gives better results when fine-tuning on 3 target tasks than training from scratch;
- we show that our 3D approach is better suited than 2D models, even when pre-trained with ImageNet;
- we finally perform an ablation study showing that leveraging the auxiliary information improves the performance for all downstream tasks and different set of transformations $\mathcal{T}$ compared to SimCLR [52].

3.2.2 Method

Problem formalization

We follow the same notations as in section 3.1.2. We want to learn an encoder $f_\theta : \mathcal{X} \to \mathbb{S}^{d-1} = \mathcal{Z}$ mapping an image x to its representation z that preserves its semantic content. In contrastive learning, each training sample $x \in \mathcal{X}$ is transformed twice through $t_1, t_2 \sim \mathcal{T}$ to produce two augmented views of the same image $(v_1, v_2) \overset{\text{def}}{=} (t_1(x), t_2(x))$, where $\mathcal{T}$ is a set of predefined transformations (see section 3.1.2). We assume they are drawn from a joint distribution $p(V_1, V_2)$.

The training procedure consists in discriminating the positive pair $(v_1, v_2) \sim p(V_1, V_2)$ from the negative pair $(v_1, v_2) \sim p(V_1)p(V_2)$, using an estimator of $I(V_1, V_2)$ crafted from f_θ[6] and called InfoNCE [211]:

$$I_{NCE}(V_1, V_2) = \mathbb{E}_{(v_1^i, v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)} \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^i)}}{\frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}} \right) \tag{3.16}$$

In Eq. 3.16, all samples $(v_2^j)_{j\neq i}$ are considered *equally* different from v_1^i (i.e. sampled independently from v_1^i). However, this is hardly true with medical images since we know, for instance, that two young healthy subjects are more similar than a young and an old healthy subject (e.g. anatomically). It means that it exists an underlying auxiliary variable $Y \in \mathbb{R}^p$ (e.g. age with $p = 1$) that should explain both V_1 and V_2. We make the following (strong) conditional independence assumption about Y:

[6] f_θ is usually defined as the composition of an encoder network $e_{\theta_1}(x)$ and a projection head z_{θ_2}(e.g. multi-layer perceptron) which is discarded after training (here $\theta = \{\theta_1, \theta_2\}$)

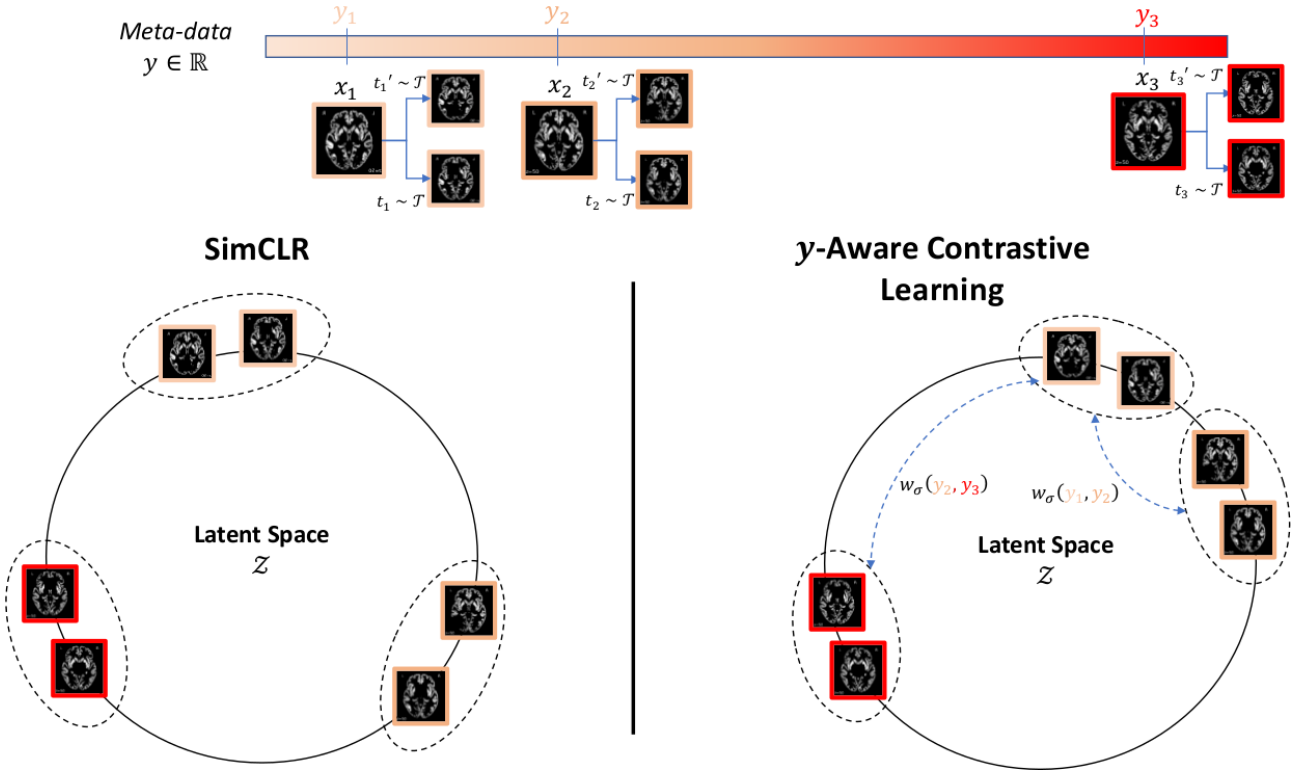


Figure 3.4: Differently from SimCLR [52], our new loss can handle auxiliary information $y \in \mathbb{R}$ by redefining the notion of similarity between two images in the latent space $\mathcal{Z}$. For an image x_i, transformed twice through two augmentations $t_1, t_2 \sim \mathcal{T}$, the resulting views $(t_1(x_i), t_2(x_i))$ are expected to be close in the latent space through the learnt mapping f_θ, as in SimCLR. However, we also expect a different input $x_{k \neq i}$ to be close to x_i in $\mathcal{Z}$ if the two *auxiliary information* y_i and y_k are similar. We define a similarity function $w_\sigma(y_i, y_k)$ that quantifies this notion of similarity.

Assumption 1. *(Conditional independence) The two views V_1 and V_2 are independent conditionally to the auxiliary variable Y, i.e: $p(V_1, V_2|Y) = p(V_1|Y)p(V_2|Y)$.*

Interpretation. The auxiliary information Y must be rich enough to carry all semantic information about V_1 and V_2. In [243], Y is called a "latent class" and is usually not observed. In Supervised Contrastive Learning [170], Y represents the (observable) label and we show here that it is a particular case in our framework. In both cases, Y was always considered as a discrete variable while here Y can be a multi-dimensional variable (with both continuous or discrete components).

Assuming 1, we can re-express the joint distribution $p(V_1, V_2)$ with the following *positive* distribution:

$$p(V_1, V_2) = \mathbb{E}_{y\sim p(Y)}\left[p(V_1|y)p(V_2|y)\right] \tag{3.17}$$

Practical issue. In a practical scenario, drawing a pair $(v_1, v_2) \sim p(V_1, V_2)$ means that we need 1) to sample the auxiliary variable $y \sim p(Y)$ and 2) sample a view $v_1 \sim p(V_1|y)$ and another (independent) view $v_2 \sim p(V_2|y)$. One important issue appears when the training set is finite with a very limited number of original images x with the same auxiliary information y. In that case, the estimation of $p(V_2|y)$ is quite poor because it relies essentially on the augmentations $\mathcal{T}$ and not on the relationship between distinct original images x sharing the same auxiliary information y.

Our proposal. We rely on *kernel density estimation* (KDE) as a workaround to estimate the conditional distribution $p(V_2|Y)$. Let $(v_2^i, y_i)_{i\in[1..N]} \overset{\text{iid}}{\sim} p(V_2, Y)$. We assume that $Y \in \mathbb{R}$ ($p = 1$) but we extend it to the multivariate case hereafter. We estimate the probability density function $p(v_2|y)$ for a pair (v_2, y) using its kernel density estimator (inspired from Nadaraya-Watson estimator [204, 301]):

$$\hat{p}(v_2|y) = \frac{\sum_{i=1}^N w_\sigma(y, y_i)\delta_{v_2^i}(v_2)}{\sum_{i=1}^N w_\sigma(y, y_i)} \tag{3.18}$$

where $w_\sigma(y, y_i) = \frac{1}{\sigma}K\left(\frac{y-y_i}{\sigma}\right)$ with K a kernel (i.e. non-negative real symmetric integrable function) and σ a bandwidth hyper-parameter to fix. Now, let $(v_1^i, v_2^i, y_i)_{i\in[1..N]} \overset{\text{iid}}{\sim} p(V_1, V_2, Y)$. Plugging the kernel density estimator into the original InfoNCE estimator leads to the following:

$$\begin{cases} \hat{p}(v_1, v_2) = \frac{1}{N}\sum_{i=1}^N \hat{p}(v_1|y_i)\hat{p}(v_2|y_i) = \frac{1}{N}\sum_{i,k=1}^N \frac{w_\sigma(y_i,y_k)}{\sum_{j=1}^N w_\sigma(y_i,y_j)}\delta_{v_1^i}(v_1)\delta_{v_2^k}(v_2) \\ I^y_{NCE}(V_1, V_2) = \frac{1}{N}\sum_{i,k=1}^N \frac{w_\sigma(y_i,y_k)}{\sum_{j=1}^N w_\sigma(y_i,y_j)} \log \frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}}{\frac{1}{N}\sum_{j=1}^N e^{f_\theta(v_1^i)\cdot f_\theta(v_2^j)}} \end{cases} \tag{3.19}$$

Where $\hat{p}(v_1|y) = \frac{1}{|C_y|}\sum_{i=1}^N \delta_{y_i}(y)\delta_{v_1^i}(v_1)$ ($C_y = \{i|y_i = y\}$) is the empirical density estimator and we assumed that $\forall i \neq j, y_i \neq y_j$ (which is almost surely true if $Y \in \mathbb{R}$). We call this estimator the ***y*-Aware InfoNCE** estimator and it is also an estimator of $I(V_1, V_2)$ under the

assumption 1. We can analyse theoretically this new estimator using the well-known kernel density estimator theory. As before, we derive the y-Aware InfoNCE loss to optimize as:

$$\mathcal{L}^{y}_{InfoNCE} = -I^{y}_{NCE}(V_1, V_2) \tag{3.20}$$

Choice of kernel. In our empirical study, we use the Gaussian kernel $K(u) \propto \exp(-\frac{u^2}{2})$ but other kernels could be explored and it is left for future work (e.g. Epanechnikov kernel).

Analysis of kernel bandwidth

Discrete case. If $(y_i)_{i\in[1..N]}$ are discrete ($Y \in \mathbb{N}$), then we cannot hypothesize that $\forall i \neq j, y_i \neq y_j$. We approximate $p(v_1, v_2)$ with empirical density estimator for both $\hat{p}(v_1|y)$ and $\hat{p}(v_2|y)$ (i.e. equivalent to impose a Delta kernel $K(u) = \delta(u)$ in previous kernel density estimator). In that case, we have:

$$\begin{cases} \hat{p}(v_1, v_2) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|C_{y_i}|^2} \sum_{k_1,k_2 \in C_{y_i}} \delta_{v_1^{k_1}}(v_1)\delta_{v_2^{k_2}}(v_2) \\ I^{y}_{NCE}(V_1, V_2) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|C_{y_i}|^2} \sum_{k_1,k_2 \in C_{y_i}} \log \frac{e^{f_\theta(v_1^{k_1})\cdot f_\theta(v_2^{k_2})}}{\frac{1}{N}\sum_{j=1}^{N} e^{f_\theta(v_1^{k_1})\cdot f_\theta(v_2^{j})}} \end{cases} \tag{3.21}$$

We retrieve the Supervised Contrastive Loss (SupCon) [170] (see Appendix B.2 for a proof) . This notably gives a new theoretical interpretation to SupCon and it relates it to information theory. It provides a first proof that SupCon optimizes an estimator of $I(V_1, V_2)$ under assumption 1. From this point-of-view, we may see $\mathcal{L}^{y}_{NCE}$ as an extension of SupCon in the continuous (and more broadly multi-dimensional) case. However, our purpose here is not to perform supervised learning but rather to build a robust encoder that can leverage auxiliary information to learn a generalizable representation of the data.

Risk and optimal bandwidth. From density estimation theory, we know that our previous density estimator has, under Lipschitz continuity and finite variance assumption, the following L_2 risk (see [125] Theorem 5.2 for detailed assumptions and a proof):

$$\mathcal{R}(\sigma) = \mathbb{E}||\hat{p}(v_2|y) - p(v_2|y)||^2_{L_2} = O\left(\frac{1}{N\sigma} + \sigma^2\right) \tag{3.22}$$

Where the $O(\cdot)$ hides a constant depending on the kernel K, $p(y)$ and first and second derivatives of $p(v_2|y)$. In practice, we use cross-validation to fix the bandwidth σ according to the performance on the downstream tasks.

Extension to multivariate auxiliary variable

If $Y \in \mathbb{R}^p$ (with $p \geq 1$) then we can extend the previous kernel density estimator to multivariate density estimator by modifying w_σ with:

$$w_\Sigma(y, y_i) = |\Sigma|^{-1/2} K(\Sigma^{-1/2}(y - y_i)) \tag{3.23}$$

here $\Sigma \in \mathbb{R}^{p \times p}$ is the bandwidth matrix (that is symmetric positive definite) and K is a symmetric kernel. As previously, we can use $K(u) \propto \exp(-\frac{uu^T}{2})$. The bandwidth matrix Σ is a hyper-parameter that needs to be fixed (e.g. through cross-validation). In particular, the correlations between auxiliary variables (i.e. non-diagonal terms in Σ) could be computed *a priori* using the training set.

Choice of the transformations $\mathcal{T}$

In our formulation, we did not specify particular transformations $\mathcal{T}$ to generate views. While there have been recent works [49, 276] proposing transformations on natural images (color distorsion, cropping, cutout [76], etc.), there is currently no consensus for medical images in the context of contrastive learning. Here, we design three sets of transformations that preserve the semantic information in MR images: cutout, random cropping and a combination of the two with also gaussian noise, gaussian blur and flip. Importantly, while color distortion is crucial on natural images [52] to avoid the model using a shortcut during training based on the color histogram, it is not necessarily the case for MR images (see Supp. 3).

3.2.3 Experiments

Datasets

We perform the experiments using a subset of BHB-10K presented in Chapter 2. In particular, we use $n = 10$k of healthy controls (HC) to perform pre-training with **participant's age as auxiliary information** Y. We make this choice because i) we know that age is rather specific to each participant and it drives the general variability ii) it is a phenotyping feature easily accessible across cohorts.

BHB-10K (subset) We aggregated 13 publicly available datasets of 3D T1 MRI scans of healthy controls (HC) acquired on more than 70 different scanners worldwide and comprising $n = 10^4$ samples. We use this dataset only to pre-train our model with the **participant's age as auxiliary information**. It corresponds to a subset of the previous dataset used in Chapter 2 (section 2.2)) where we use OpenBHB along with HCP [286], OASIS 3 [184] and ICBM [198].

Then, we study several real-world clinical problems with patients suffering from schizophrenia (SCZ), bipolar disorder (BD) and Alzheimer's disease (AD), thus covering both psychiatric and neurological disorders. Specifically, the learned representation is tested using the following clinical data-sets (as in Chapter 2 excepted for ADNI):

- **SCHIZCONNECT-VIP**[7] It comprises $n = 605$ multi-site MRI scans including 275 patients with strict schizophrenia (SCZ) and 330 HC;

[7] http://schizconnect.org

- **BIOBD** [147, 241] This dataset includes $n = 662$ MRI scans acquired on 8 different sites with 356 HC and 306 patients with bipolar disorder (BD);
- **BSNIP** [271] It includes $n = 511$ MRI scans with $n = 200$ HC, $n = 194$ SCZ and $n = 117$ BD. This independent dataset is used only at test time in Fig. 3.5b);
- **Alzheimer's Disease Neuroimaging Initiative (ADNI-GO)**[8] We use $n = 387$ co-registered T1-weighted MRI images divided in $n = 199$ healthy controls and $n = 188$ Alzheimer's patients (AD). Since it is a longitudinal study and we did not want to bias our analysis, we only included one scan per patient at the first session (baseline) and we performed a visual quality check. Furthermore, all patients included have a constant follow-up clinical status (either control or AD).

All these data-sets have been pre-processed in the same way with a non-linear registration to the MNI template and a gray matter extraction step. The final spatial resolution is $1.5mm$ isotropic and the images are of size $121 \times 145 \times 121$.

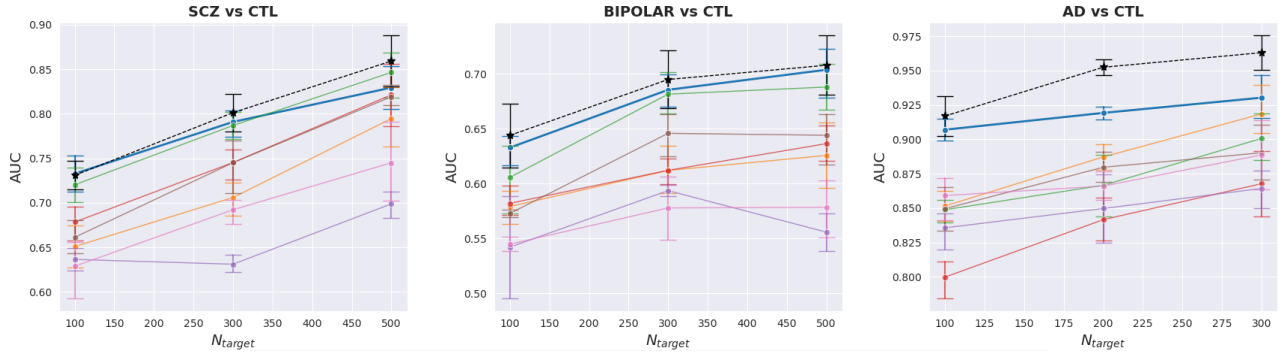
Implementation details

We implement our new loss based on the original InfoNCE loss [52] with Pytorch [215] and we use the Adam optimizer during training. As opposed to SimCLR [52] and in line with a recent study on medical imaging [47], we only use a batch size of $N = 64$ as it did not significantly change our results (see Results section). We also follow [52] by fixing $\tau = 0.1$ in Eq. 3.16 and Eq. 3.20 and we set the learning rate to $\alpha = 10^{-4}$, decreasing it by 0.9 every 10 epochs. During pre-training, we use an encoder $f_\theta = z_{\theta_2} \circ e_{\theta_1}$ with e_{θ_1} a 3D adaptation of DenseNet121 [149] and z_{θ_2} a projection head (a 2-layers MLP as in [52]). This projection head is discarded for fine-tuning/evaluating the representation. Our code is publicly available here: https://github.com/Duplums/yAwareContrastiveLearning

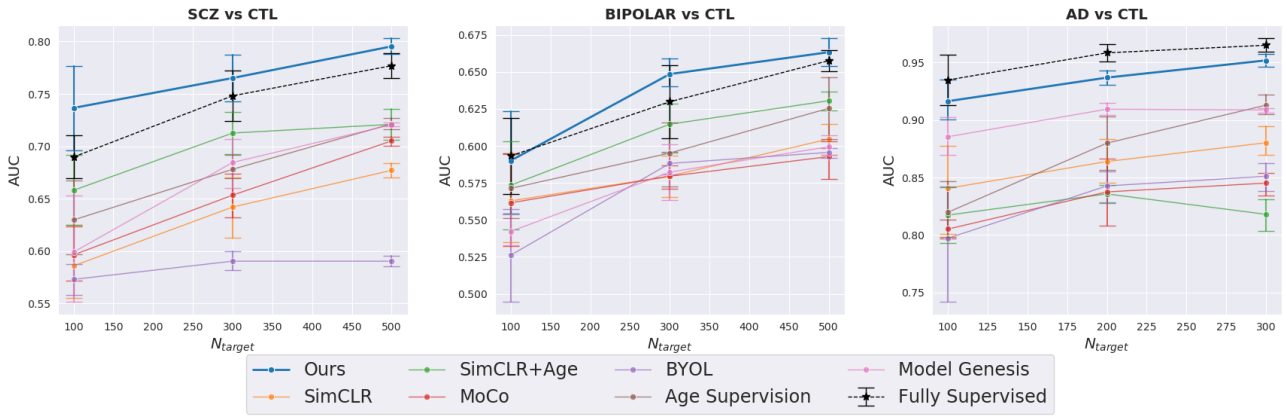
Evaluation of the representation

In Fig.3.5, we compare the representation learned using our model f_θ with the ones estimated using i) the original InfoNCE loss (a.k.a SimCLR) [52], ii) Model Genesis [320], a SOTA model for self-supervised learning with medical images using context-restoration, iii) a standard pre-training on age using a supervised approach (i.e. l_1 loss for age prediction), iv) BYOL [120] and MoCo [136] (memory bank $K = 1024$) , two recently proposed SOTA models for representation learning, v) a multi-task approach SimCLR with age regression in the latent space (SimCLR+Age) and a fully fine-tuned supervised DenseNet trained to predict the final task. This can be considered as an upper bound, if the training data-set were sufficiently large (*e.g.,* ImageNet). Nevertheless, in our case it may be outperformed when images in downstream tasks are hard to classify and only a few are accessible.

[8] http://adni.loni.usc.edu/about/adni-go



(a) 5-fold CV Stratified on Site



(b) 5-fold CV Leave-Site-Out

Figure 3.5: Comparison of different representations in terms of classification accuracy (downstream task) on three different data-sets (one per column). Classification is performed using a linear layer on top of the pre-trained frozen encoders. (a) Data for training/validation and test come from the the same acquisition sites (b) Data for training/validation and test come from different sites.

For the pre-training of our algorithm f_θ, we only use the BHB dataset with the participant's age as auxiliary information. For both contrastive learning methods and BYOL, we fix $\sigma = 5$ in Eq. 3.20 and Eq. 3.16 and only use random cutout for the transformations $\mathcal{T}$ with a black patch covering $p = 25\%$ of the input image. We use UNet for pre-training with Model Genesis and DenseNet121 for all other models.

In order to evaluate the quality of the learnt representations, we only added a linear layer on top of the frozen encoders pre-trained on BHB. We tune this linear layer on 3 different binary classification tasks (see Datasets section) with 5-fold cross-validation (CV). We tested two different situations: data for training/validation and test come either from the same sites (first row) or from different sites (second row). We also vary the size (i.e. number of subjects, N_{target}) of the training/validation set. For (a), we perform a stratified nested CV (two 5-fold CV, the inner one for choosing the best hyper-parameters and the outer one for estimating the test error). For (b), we use a 5-fold CV for estimating the best hyper-parameters and keep an independent constant test set for all N_{target}.

From Fig. 3.5, we notice that our method consistently outperforms the other pre-trainings

even in the very small data regime ($N = 100$) and it matches the performance of the fully-supervised setting on 2 data-sets. Differently from age supervision, f_θ is less specialized on a particular proxy task and it can be directly transferred on the final task at hand without fine-tuning the whole network. Furthermore, compared to the multi-task approach SimCLR+Age, the features extracted by our method are less sensitive to the site where the MR images are coming from. This shows that our technique is the only one that efficiently uses the highly multi-centric dataset BHB by making the features learnt during pre-training less correlated to the acquisition sites.

Importance of σ and $\mathcal{T}$ in the positive sampling

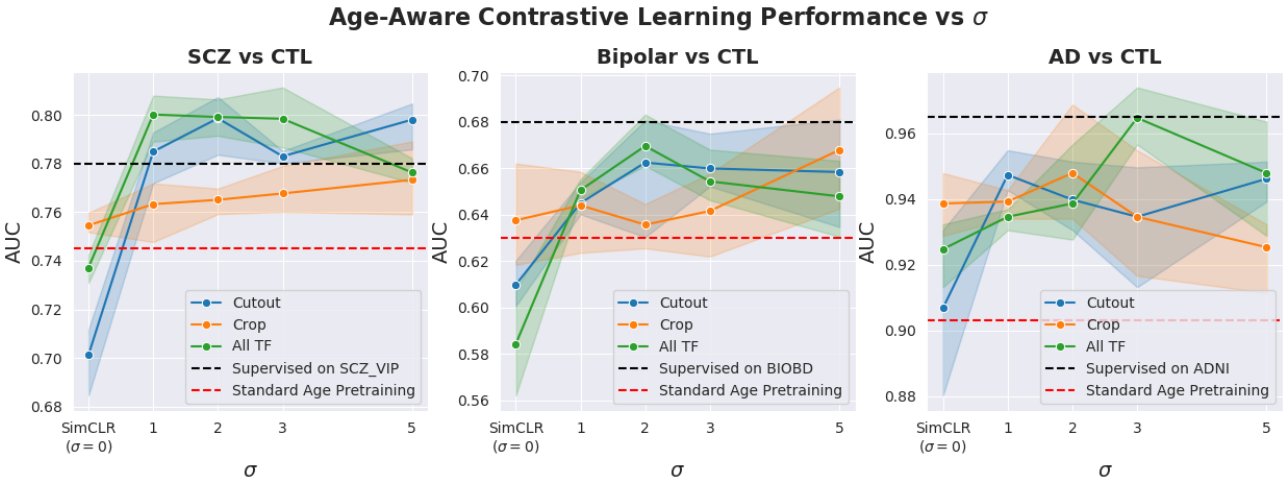


Figure 3.6: Linear classification performance on three binary classification tasks with $N_{pretrained} = 10^4$. *All TF* includes crop, cutout, gaussian noise, gaussian blur and flip. The encoder is frozen and we only tune a linear layer on top of it. $\sigma = 0$ corresponds to SimCLR [52] with InfoNCE loss. As we increase σ, we add more positive examples for a given anchor x_i with close auxiliary information (i.e. close age here).

In Fig. 3.6, we study the impact of σ in Eq. 3.20 on the final representation learnt for a given set of transformations $\mathcal{T}$. As highlighted in [52], hard transformations seem to be important for contrastive learning (at least on natural images), therefore we have evaluated three different sets of transformations $\mathcal{T}_1 = \{$ Random Crop $\}$, $\mathcal{T}_2 = \{$ Random Cutout $\}$ and $\mathcal{T}_3 = \{$ Cutout, Crop, Gaussian Noise, Gaussian Blur, Flip $\}$. Importantly, we did not include color distorsion in $\mathcal{T}_3$ since i) it is not adapted to MRI images where a voxel's intensity encodes a gray matter density and ii) we did not observe significant difference between the color histograms of different scans as opposed to [52] (see next section). As before, we evaluated our representation under the linear evaluation protocol. We can observe that $\mathcal{T}_2$ and $\mathcal{T}_3$ give similar performances with $\sigma > 0$, always outperforming both SimCLR ($\sigma = 0$) and age supervision on BHB. It also even outperforms the fully-supervised baseline on SCZ vs HC. We also find that a strong cropping or cutout strategy is detrimental for the final performances (see next section). Since $\mathcal{T}_2$ is computationally less expensive than $\mathcal{T}_3$, we chose to use $\mathcal{T} = \mathcal{T}_2$ and $\sigma = 5$ in our experiments.

Transfer learning results

Next, we fine-tune the whole encoder f_θ with different initialization on the 3 downstream tasks (see Table 3.1). To be comparable with Model Genesis [320], we also use the same UNet backbone for f_θ and we still fixed $\mathcal{T}_2 = \{\text{Random Cutout}\}$ and $\sigma = 5$. First, our approach outperforms the CNNs trained from scratch on all tasks as well as Model Genesis, even with the same backbone. Second, when using DenseNet, our pre-training remains better than using age supervision as pre-training for SCZ vs HC (even with the same transformations) and it is competitive on BD vs HC and AD vs HC.

Backbone	Pre-training	SCZ vs HC		BD vs HC		AD vs HC	
		$N_{train} = 100$	$N_{train} = 500$	$N_{train} = 100$	$N_{train} = 500$	$N_{train} = 100$	$N_{train} = 300$
UNet	None	$72.62_{\pm 0.9}$	$76.45_{\pm 2.2}$	$63.03_{\pm 2.7}$	$69.20_{\pm 3.7}$	$88.12_{\pm 3.2}$	$94.16_{\pm 3.9}$
	Model Genesis [320]	$73.00_{\pm 3.4}$	$81.8_{\pm 4.7}$	$60.96_{\pm 1.8}$	$67.04_{\pm 4.4}$	$89.44_{\pm 2.6}$	$95.16_{\pm 3.3}$
	SimCLR [49]	$73.63_{\pm 2.4}$	$80.12_{\pm 4.9}$	$59.89_{\pm 2.6}$	$66.51_{\pm 4.3}$	$90.60_{\pm 2.5}$	$94.21_{\pm 2.7}$
	Age Prediction w/ D.A	$\underline{75.32}_{\pm 2.2}$	$\underline{85.27}_{\pm 2.3}$	$\mathbf{64.6}_{\pm \mathbf{1.6}}$	$\mathbf{70.78}_{\pm \mathbf{2.1}}$	$\underline{91.71}_{\pm 1.1}$	$\underline{95.26}_{\pm 1.5}$
	Age-Aware Contrastive Learning (ours)	$\mathbf{75.95}_{\pm 2.7}$	$\mathbf{85.73}_{\pm 4.7}$	$\underline{63.79}_{\pm 3.0}$	$\underline{70.35}_{\pm 2.7}$	$\mathbf{92.19}_{\pm \mathbf{1.8}}$	$\mathbf{96.58}_{\pm \mathbf{1.6}}$
DenseNet	None	$73.09_{\pm 1.6}$	$85.92_{\pm 2.8}$	$64.39_{\pm 2.9}$	$70.77_{\pm 2.7}$	$92.23_{\pm 1.6}$	$93.68_{\pm 1.7}$
	None w/ D.A	$\underline{74.71}_{\pm 1.3}$	$86.94_{\pm 2.8}$	$64.79_{\pm 1.3}$	$72.25_{\pm 1.5}$	$92.10_{\pm 1.8}$	$94.16_{\pm 2.5}$
	SimCLR [52]	$70.80_{\pm 1.9}$	$86.35_{\pm 2.2}$	$60.57_{\pm 1.9}$	$67.99_{\pm 3.3}$	$91.54_{\pm 1.9}$	$94.26_{\pm 2.9}$
	BYOL [120]	$69.55_{\pm 2.4}$	$82.73_{\pm 2.2}$	$58.94_{\pm 3.8}$	$66.34_{\pm 3.7}$	$90.19_{\pm 2.0}$	$90.0_{\pm 3.7}$
	MoCov2 [136]	$72.02_{\pm 0.03}$	$82.48_{\pm 3.9}$	$60.29_{\pm 2.4}$	$68.77_{\pm 4.0}$	$87.0_{\pm 2.9}$	$91.31_{\pm 3.8}$
	Age Prediction	$72.90_{\pm 4.6}$	$\underline{87.75}_{\pm 2.0}$	$64.60_{\pm 3.6}$	$72.07_{\pm 3.0}$	$92.07_{\pm 2.7}$	$\underline{96.37}_{\pm 0.9}$
	Age Prediction w/ D.A	$74.06_{\pm 3.4}$	$86.90_{\pm 1.6}$	$\mathbf{65.79}_{\pm 2.0}$	$\underline{73.02}_{\pm 4.3}$	$\mathbf{94.01}_{\pm \mathbf{1.4}}$	$96.10_{\pm 3.0}$
	Age-Aware Contrastive Learning (ours)	$\mathbf{76.33}_{\pm 2.3}$	$\mathbf{88.11}_{\pm 1.5}$	$\underline{65.36}_{\pm 3.7}$	$\mathbf{73.33}_{\pm \mathbf{4.3}}$	$\underline{93.87}_{\pm 1.3}$	$\mathbf{96.84}_{\pm \mathbf{2.3}}$

Table 3.1: Fine-tuning results using $N_{train} = 100$ and $N_{train} = 500$ ($N_{train} = 300$ for AD vs HC) training subjects. For each task, we report the AUC (%) of the fine-tuned models initialized with different approaches with 5-fold cross-validation. We use $\sigma = 5$ for the Age-Aware InfoNCE loss. For age prediction, we employ the same transformations as in contrastive learning for the Data Augmentation (D.A) strategy. Only the encoder of UNet is used when fine-tuning on the downstream tasks. Best results are in **bold** and second bests are underlined.

Comparison with linear models

In the previous chapter, we have demonstrated that CNN performed on par with regularized linear models at current samples size ($N_{train} \approx 1$k) – at least for the detection of psychiatric disorders (in particular schizophrenia, bipolar disorder and autism). The integration of epistemic uncertainty with deep ensemble allowed to improve significantly classifiers performance and calibration. Considering the previous improvement with our new transfer learning strategy, we ask whether i) the proposed transfer learning strategy induces better generalization performance than linear models and ii) we can combine deep ensemble learning with transfer learning to outperform all previous approaches. We take the same experimental design than in Chapter 2 (see section 2.2) to answer.

As before, we take the pre-trained DenseNet121 network with $\mathcal{T}_2$ transformations and $\sigma = 5$ and we fine-tune all weights on the same three target tasks as in Chapter 2: SCZ vs HC, BD vs HC and ASD vs HC. Differently from the previous TL experiments, we consider much more training samples on these tasks ($\approx 2\times$ and $1.6\times$ more resp. for SCZ vs HC and BD vs HC) and we evaluate the model only on psychiatric disorders classification.

Task	Test Set	Deep Learning Models				SML		
		Baseline	Deep Ensemble	Transfer	Transfer + Deep Ensemble	rbf-SVM	Logistic ℓ_2	ElasticNet
SCZ vs. HC ↑	Internal Test	$85.27_{\pm 1.60}$	$85.73_{\pm 0.53}$	$85.17_{\pm 0.37}$	$\mathbf{86.28}_{\pm \mathbf{0.44}}$ (+1.01)	$82.06_{\pm 0.00}$	$84.03_{\pm 0.00}$	$85.98_{\pm 1.9}$
$N_{train} = 933$	External Test	$75.52_{\pm 0.12}$	$\mathbf{77.47}_{\pm \mathbf{0.71}}$	$77.00_{\pm 0.55}$	$76.36_{\pm 0.61}$ (+0.84)	$72.88_{\pm 0.95}$	$73.60_{\pm 0.00}$	$76.42_{\pm 1.68}$
BD vs. HC ↑	Internal Test	$76.49_{\pm 2.16}$	$79.49_{\pm 1.36}$	$78.81_{\pm 2.48}$	$\mathbf{79.59}_{\pm \mathbf{1.77}}$ (+3.10)	$73.63_{\pm 0.00}$	$72.96_{\pm 0.25}$	$73.85_{\pm 0.28}$
$N_{train} = 832$	External Test	$68.57_{\pm 4.72}$	$76.11_{\pm 0.53}$	$77.06_{\pm 1.90}$	$\mathbf{78.01}_{\pm \mathbf{1.97}}$ (+9.44)	$63.92_{\pm 0.00}$	$70.12_{\pm 0.26}$	$70.26_{\pm 1.75}$
ASD vs. HC ↑	Internal Test	$65.74_{\pm 1.47}$	$67.67_{\pm 0.74}$	$66.36_{\pm 1.14}$	$\mathbf{68.48}_{\pm \mathbf{1.45}}$ (+2.74)	$66.84_{\pm 0.00}$	$63.40_{\pm 0.18}$	$60.62_{\pm 2.63}$
$N_{train} = 1526$	External Test	$62.93_{\pm 2.40}$	$64.48_{\pm 1.51}$	$68.76_{\pm 1.70}$	$\mathbf{69.68}_{\pm \mathbf{1.70}}$ (+6.75)	$60.28_{\pm 0.00}$	$61.85_{\pm 0.05}$	$54.96_{\pm 4.94}$

Table 3.2: Deep Ensemble learning and Transfer Learning from a healthy dataset largely improve DL performance over SML models, especially on complex tasks such as ASD and BD detection. We report average AUC for all models and the standard deviation by repeating each experiment three times. For all DL results, we use DenseNet121 as backbone. The baseline corresponds to a single network trained from scratch on VBM images. For Deep Ensemble, we aggregate three networks trained from different random initialization. For Transfer, we pre-traine a single network with Age-Aware contrastive learning and we fine-tune it on each clinical task. For Transfer+Deep Ensemble, we aggregate three networks, all pre-trained with Age-Aware contrastive learning (only once) and fine-tune on each downstream task. The randomness thus comes from the gradient descent optimization on each downstream task. Green numbers indicate improvement over DL baselines.

From Table 3.2, we observe a consistent increase in performance when combining both Deep Ensemble learning and Transfer Learning w.r.t. baseline on the external test (+0.84%, +9.44%, +6.75% AUC resp. on schizophrenia, bipolar disorder, and autism spectrum disorders detection). For Deep Ensemble learning, it supports the hypothesis that different random initialization leads to different representations after training. For Transfer Learning, it shows that anatomical features learnt from the healthy population during brain maturation and aging can be re-used, in particular to drastically improve DL generalization performance on the external test for hard clinical tasks (i.e bipolar disorder and autism spectrum disorders). Nonetheless, DL performance is still on par with SML models on easier tasks (e.g., schizophrenia), the task difficulty being measured by linear performance.

These findings suggest that i) discriminative transferable anatomical non-linear patterns can be learned with DL through pre-training from brain imaging of the healthy population; ii) different DL initialization converge to different solutions after training that, if aggregated together, can outperform SML; iii) DL models tend to learn simple features on easy tasks (such as schizophrenia detection), falling into the Simplicity Bias [253], which encourages CNN to find the simplest features to perform the task (and thus hurting generalization power on external test sets).

2D vs 3D approach and transfer learning from ImageNet

Previous models have been trained directly on 3D volumes, by extending 2D to 3D kernels in CNN architectures. Another common strategy in medial image analysis is to see each 3D volume as a collection of 2D scans and to perform prediction using a 2D CNN pre-trained on ImageNet. This approach does not account for the 3D spatial structure of brain mages and it also assumes independence between 2D scans of to the same 3D brain volume. For completeness, we evaluate this strategy on our datasets using 2 backbones: ResNet18 and DenseNet121. We

Backbone	Pre-training	SCZ vs HC		BD vs HC		AD vs HC	
		$N_{train} = 100$	$N_{train} = 500$	$N_{train} = 100$	$N_{train} = 500$	$N_{train} = 100$	$N_{train} = 300$
2D-ResNet18	None	$73.55_{\pm 0.74}$	$82.72_{\pm 2.27}$	$61.66_{\pm 2.06}$	$69.51_{\pm 4.05}$	$91.84_{\pm 0.93}$	$95.00_{\pm 1.23}$
	ImageNet	$73.85_{\pm 2.22}$	$85.89_{\pm 1.99}$	$64.60_{\pm 2.06}$	$70.79_{\pm 1.75}$	$91.78_{\pm 1.22}$	$94.25_{\pm 1.62}$
2D-DenseNet	None	$74.17_{\pm 2.32}$	$82.19_{\pm 3.82}$	$64.12_{\pm 2.45}$	$69.84_{\pm 4.22}$	$89.40_{\pm 1.25}$	$92.92_{\pm 2.17}$
	ImageNet	$73.93_{\pm 1.71}$	$84.06_{\pm 2.88}$	$64.45_{\pm 1.94}$	$71.83_{\pm 2.96}$	$91.07_{\pm 1.07}$	$94.66_{\pm 1.60}$
3D-DenseNet	None	$73.09_{\pm 1.6}$	$85.92_{\pm 2.8}$	$64.39_{\pm 2.9}$	$70.77_{\pm 2.7}$	$92.23_{\pm 1.6}$	$93.68_{\pm 1.7}$
	Age-Aware CL	$\mathbf{76.33}_{\pm 2.3}$	$\mathbf{88.11}_{\pm 1.5}$	$\mathbf{65.36}_{\pm 3.7}$	$\mathbf{73.33}_{\pm \mathbf{4.3}}$	$\mathbf{93.87}_{\pm 1.3}$	$\mathbf{96.84}_{\pm \mathbf{2.3}}$

Table 3.3: Fine-tuning results using a 2D approach for brain MRI classification. We represent each 3D volume as a collection of 2D scans along the axial plane, following [26]. At test time, we use the median prediction for all 2D scans of the same volume and we report AUC(%). As before, we use a 5-fold CV and set $\sigma = 5$ for Age-Aware InfoNCE loss. Best results are in **bold** and second bests are underlined.

expand each 3D volume along the axial plane and we retain only the central 70 slices, following the experimental setup in [26] that studied transfer learning for psychiatric disorder prediction using exclusively a 2D approach.

In Table 3.3, we observe that our 3D pre-training always gives the best results compared to all 2D approaches and backbones. In more details, ImageNet pre-training improves performance consistently over random initialization using a 2D approach only when $N > 300$. Additionally, ImageNet pre-training gives comparable results with 3D models trained from scratch. We also observe that, without pre-training, 2D models give always worse or comparable performance than their 3D counterpart. Overall, these results suggest that our 3D approach is well adapted, as it may account for 3D spatial structure of brain images.

Remark. Our approach is best suited to 3D volumes than 2D images. Indeed, auxiliary information y is the same for all slices of the same brain MRI so y-Aware InfoNCE imposes equal constraints for all of them in the latent space. We argue this is sub-optimal since each slice brings different anatomical information (*e.g.,* slices in the parietal and temporal lobes for 3D volumes cut along axial plane). In other words, we cannot make independence hypothesis between several slices of the same brain MRI. Consequently, an additional pairing strategy is required to impose constraints only for anatomically similar 2D slices across subjects (such as [182]). We have left it for future work since the 2D approach performs worse than its 3D counterpart (possibly because it does not account for the original 3D spatial structure of the brain).

Visualization of latent space

We qualitatively show that our model encourages images with close auxiliary information Y to be close in the latent space by plotting the 2D UMAP representation of the encoded images from ADNI data-set (unseen during training). In Fig. 3.7, we make two observations: 1) there is a continuum in images representation with our model according to chronological age (suggesting that our encoder captures biological variability from MRI) and 2) pathological

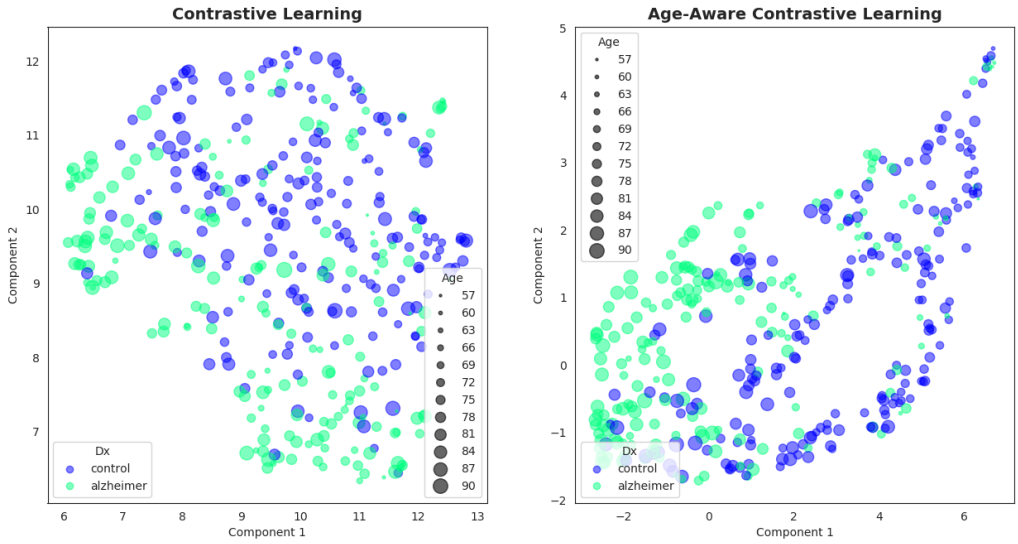


Figure 3.7: 2D UMAP of ADNI features encoded (left) with SimCLR pre-training; (right) with our method. MRI from healthy participants with approximately the same age are mapped to the same region with our model. It is also able to discriminate AD patients from HC *without* fine-tuning on the downstream task.

brains (here with Alzheimer's disease) follows a distinct trajectory in the latent space than healthy ones, suggesting that our encoder also captures pathological variability, even if it has never been explicitly trained on brain imaging with AD. It further explains the quantitative results obtained previously with linear probing, that showed close performance between our pre-trained model (without fine-tuning) and a fully supervised model trained to predict clinical status.

Influence of batch size and data augmentation strength

As pointed out in SimCLR [52], batch size is a critical hyper-parameter when performing contrastive learning (at least on vision tasks). In this work, we tested 2 batch size $N \in \{64, 100\}$ and we assessed the quality of the representation with a linear probe on the previous downstream tasks.

Batch Size	**Target Task**		
	SCZ vs HC	BIP vs HC	AD vs HC
64	$82.94_{\pm 2.7}$	$70.36_{\pm 2.6}$	$93.03_{\pm 1.8}$
100	$84.15_{\pm 2.7}$	$70.42_{\pm 1.1}$	$93.53_{\pm 1.6}$

Table 3.4: AUC score (%) as we vary the batch size during pre-training.

In Fig. 3.4, we do not observe significant difference for bigger batch size, in line with [47], studying segmentation of medical images with contrastive learning. It suggests that **a large batch size is not required when dealing with medical images**. We hypothesize that the mutual information (MI) $I(V_1, V_2)$ is not very high for brain imaging (typically $\leq \log N$) so the InfoNCE estimator can well approximate it. In other words, InfoNCE is less biased on brain

images than natural images. It suggests that 2 views are visually more similar with natural images than medical ones, thus leading to a higher MI for the former than the latter.

Next, we perform an ablation study on the augmentation strength required for our images. Specifically, we vary crops size and cutout size (i.e. size of black covering patches) and we report AUC under linear evaluation. We pre-train our model using $\sigma = 5$ and we set $N_{target} = 500$ for SCZ vs HC and $N_{target} = 300$ for AD vs HC. In Fig. 3.8, we observe that a strong augmentation strategy is not as critical as for original SimCLR on natural images.

Transformations		**Target Task**		
		SCZ vs HC	BIP vs HC	AD vs HC
Cutout	$p = 25\%$	$82.94_{\pm 2.7}$	$70.36_{\pm 2.6}$	$93.03_{\pm 1.8}$
	$p = 50\%$	$84.00_{\pm 2.1}$	$68.96_{\pm 2.2}$	$89.21_{\pm 2.7}$
Crop	$p' = 75\%$	$84.73_{\pm 0.7}$	$69.77_{\pm 4.3}$	$94.88_{\pm 2.7}$
	$p' = 50\%$	$81.77_{\pm 3.1}$	$68.69_{\pm 1.3}$	$91.46_{\pm 3.2}$

Figure 3.8: AUC score (%) over 3 different downstream tasks with $N_{target} = 500$ for SCZ vs HC and BIP vs HC and $N_{target} = 300$ for AD vs HC. The black patch size p (for random cutout) and the crop size p' are set during the pre-training in the contrastive learning framework and we fixed $\sigma = 5$. We only tune a linear probe on top of the pre-trained encoder and we perform a 5-fold cross validation. Based on these results, we fixed $p = 25\%$ and $p' = 75\%$ in this study.

Possible bias with color histogram

Classically, in the brain MR images pre-processed with gray matter extraction and non-linear registration, the voxel intensity encodes the gray matter density in this voxel. It is intrinsically different from the natural images where a pixel encodes an RGB value. Here, it does not make sense to apply color distortion to our images. However, as noted in [52], the model may learn a shortcut during the training if we solely apply cropping or cutout. We have plotted figure 3.9 the histogram of voxel intensities for 2 different images i) randomly cropped ii) with random cutout.

Differently from SimCLR on natural images, the network should not be able to use the color histogram to perform instance discrimination, thus comforting our choice for using cutout without color distortion.

3.2.4 Conclusion

Our key contribution is the introduction of a new contrastive loss, which leverages continuous (and more broadly multi-dimensional) auxiliary information for medical images in a self-supervised setting. We showed that our model, pre-trained with a large heterogeneous brain MRI dataset ($n = 10^4$) of healthy subjects, outperforms the other SOTA methods on three binary classification tasks. In some cases, it even reaches the performance of a fully-supervised network without fine-tuning. This demonstrates that our model can learn a meaningful and relevant representation of healthy brains which can be used to discriminate patients in small

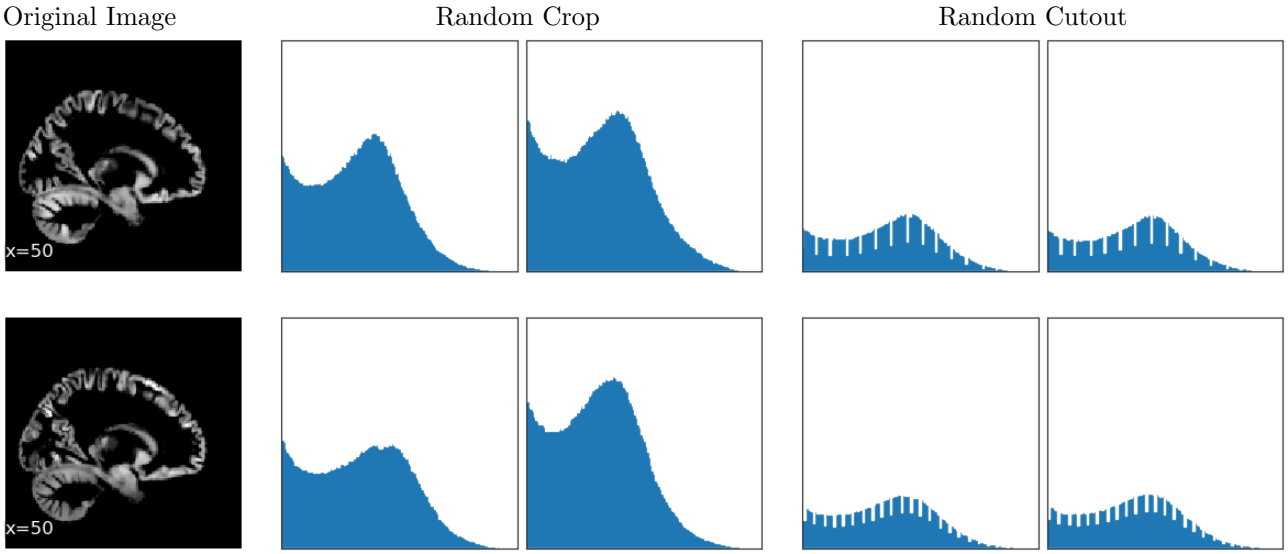


Figure 3.9: Histogram of pixel intensities for 2 different images either i) randomly cropped or ii) partially masked with random cutout. We do not observe strong differences between the histograms for a given transformation. As such, color distortion may not be as critical as in [52] to learn a robust representation since the network cannot take a shortcut based only on the color histogram.

data-sets. An ablation study showed that our method consistently improves upon SimCLR for three different sets of transformations. We also made a step towards a debiased algorithm by demonstrating that our model is less sensitive to the site effect than other SOTA fully supervised algorithms trained from scratch. We think this is still an important issue leading to strong biases in machine learning algorithms and it currently leads to costly harmonization protocols between hospitals during acquisitions. Finally, as a step towards reproducible research, we made our code public and we released the OpenBHB dataset[9] (subset of BHB-10K) to the scientific community.

Future work will consist in developing transformations more adapted to medical images in the contrastive learning framework and in integrating other available auxiliary information (*e.g* cognition) and modalities (*e.g* genetics). Finally, we envision to adapt the current framework for longitudinal studies (such as ADNI).

3.3 Theoretical analysis and prior for contrastive learning

The theory exposed previously mainly relies on information theory where we seek to estimate the mutual information (MI) between views V_1 and V_2 in order to capture the semantic content inside images. But from this perspective, can we prove that optimizing InfoNCE leads to a good representation? Is MI the right tool to explain the success of contrastive learning ?

In 2020, Tschannen et al. [281] has empirically shown that MI alone cannot explain the current success of CL. One fundamental observation was that, over a family of bijective DNN encoders $\{f_\theta\}$, some representations can be arbitrarily good or bad for a given downstream task,

[9] https://ieee-dataport.org/open-access/openbhb-multi-site-brain-mri-dataset-age-prediction-and-debiasing

Hexadecimal Representation

Human-Readable Representation

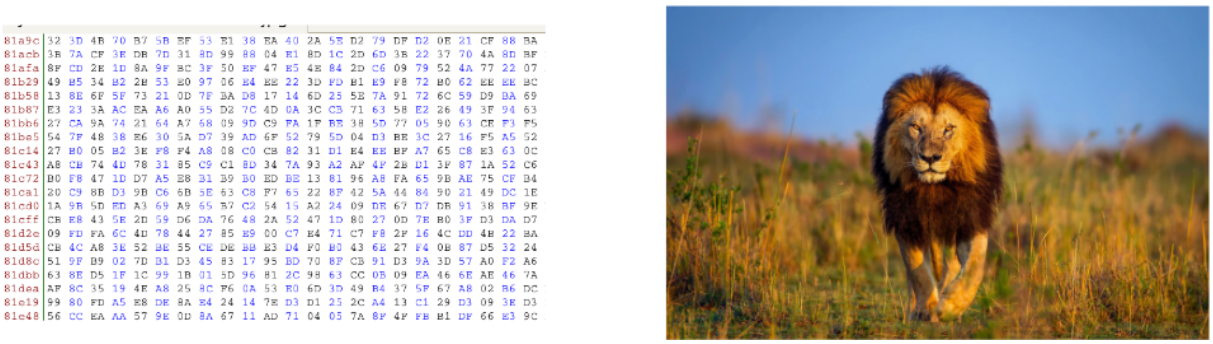


Figure 3.10: Taken from [5]. Hexadecimal representation on the left has more information content than image on the right. However, human brain only processes the latter to take immediate vital decision of whether to escape or not. Quantifying information content with entropy is not enough to characterize representation structure. Information theory does not provide a satisfactory framework to fully explain the current success of contrastive learning leading to "good" representation.

while always preserving MI: $I(V_1, V_2) = I(f_\theta(V_1), f_\theta(V_2))$ for all θ (because they are bijective). In other words, these representations contain the same amount of information (as measured by entropy) but only some of them may linearly separate semantic classes (such as objects in natural images) while others may have completely random structures. In some way, it relates to a previous observation made by Alain and Bengio [5] when introducing the concept of *linear probe*: "*neural networks are really about distilling computationally-useful representations, and they are not about information contents as described by the field of Information Theory*". It is well illustrated in Fig. 3.10 where only one representation can be efficiently processed by human brain to save its life. In our context, only measuring the information content in the DNN representation $f_\theta(V_1)$ is not enough to guarantee a useful representation for a subsequent downstream task.

As a result, in the following we turn to a metric learning perspective for CL. We first show that CL optimizes two important properties leading to a desirable representation: *alignment* between positive samples (drawn from $p(V_1, V_2)$) and *uniformity* between negative samples (drawn from $p(V_1)p(V_2)$). In particular, alignment is noticeably stronger than mere information preservation property. Then, built on this analysis, we introduce a new loss function, called *Decoupled Uniformity* that elegantly optimizes alignment and uniformity in a multi-view setting without requiring a large batch size. We theoretically analyze this loss and we prove first generalization guarantees under strong assumptions on the data augmentation strategy, which is the main bottleneck of CL limiting its wide applications across visual domains (e.g. medical imaging). Next, we ask whether *prior knowledge* can be integrated into CL to relax these assumptions while still ensuring generalization guarantees on downstream task. In a practical scenario, this prior knowledge can be of two kinds: i) given by generative models (unsupervised scenario) or ii) given as auxiliary attributes (weakly-supervised scenario as in previous section). This framework notably allows a direct connection between generative models and CL for the

first time (to the best of our knowledge).

We finally provide empirical evidence supporting our theory on standard vision benchmarks and we then apply it to real-world scenario with our brain imaging datasets.

3.3.1 Contrastive learning optimizes alignment and uniformity

We take the same notations as in previous sections 3.1.2 and 3.2. We recall the InfoNCE objective optimized in CL (see section 3.1.2):

$$\mathcal{L}_{InfoNCE} = -\mathbb{E}_{(v_1^i, v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)} \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^i)}}{\frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}} \right) \tag{3.24}$$

where $p(V_1, V_2)$ is the positive distribution, $p(V_1)p(V_2)$ is the negative distribution and $f_\theta : \mathcal{X} \to \mathbb{S}^{d-1} = \mathcal{Z}$ is the encoder. We can decompose $\mathcal{L}_{InfoNCE}$ into two terms [297] (see proof in Appendix B.3):

$$\begin{aligned} \mathcal{L}_{InfoNCE} &= -\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)} \left(f_\theta(v_1)\cdot f_\theta(v_2)\right) + \mathbb{E}_{(v_1,v_2^1)\sim p(V_1,V_2)(v_2^k)_{k\neq 1}\sim p(V_2)} \log \frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1)\cdot f_\theta(v_2^k)} \\ &\xrightarrow[N\to\infty]{} \underbrace{-\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)} \left(f_\theta(v_1)\cdot f_\theta(v_2)\right)}_{\text{Alignment}} + \underbrace{\mathbb{E}_{v_1\sim p(V_1)} \log \mathbb{E}_{v_2\sim p(V_2)} e^{f_\theta(v_1)\cdot f_\theta(v_2)}}_{\text{Uniformity}} \end{aligned} \tag{3.25}$$

This decomposition gives new insight when optimizing InfoNCE with a large batch size $N \gg 1$. Optimizing *alignment* imposes representation of two positive samples to be close while uniformity imposes uniform distribution of representations in the latent space, as we will see. For further analysis, we introduce the following 2 metrics [297]:

1. **Alignment** $\mathcal{L}_{align} = \mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)} ||f_\theta(v_1) - f_\theta(v_2)||^2 = 2 - 2\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)} f_\theta(v_1)\cdot f_\theta(v_2)$ since the latent space is a hyper-sphere (i.e. $||f_\theta(v)|| = 1$)
2. **Uniformity** $\mathcal{L}_{unif} = \log \mathbb{E}_{(v_1,v_2)\sim p(V_1)p(V_2)} \exp\left(-||f_\theta(v_1) - f_\theta(v_2)||^2\right)$ which has close connection with $\mathcal{L}_{InfoNCE}$

While the link between alignment in $\mathcal{L}_{InfoNCE}$ and $\mathcal{L}_{align}$ is obvious, it is not the case for uniformity metric $\mathcal{L}_{unif}$. Wang and Isola [297] have proven the following important result:

Theorem 1. *(Optimal Uniformity) Assuming that $p(V_1) = p(V_2)$, then any optimal minimizer θ^* of $\mathcal{L}_{unif}$ are such that $f_{\theta^*}(v)$ are uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$ for $v \sim p(V_1)$. If they exist, they are the same minimizers as for the uniformity term in InfoNCE when $N \to \infty$.*

It basically means that we can push the log outside expectation in Eq. 3.25 while preserving the same minimizers. As a result, when $N \gg 1$, optimizing InfoNCE seeks for minimizers θ^* that are i) perfectly aligned (i.e. $f_{\theta^*}(v_1) = f_{\theta^*}(v_2)$ for all positive pairs $(v_1, v_2) \sim p(V_1, V_2)$) and ii) perfectly uniformed (i.e. $\{f_\theta(v)\}$ are uniformly distributed for $v \sim p(V_1)$). The question is: can we realize both perfect alignment and uniformity ?

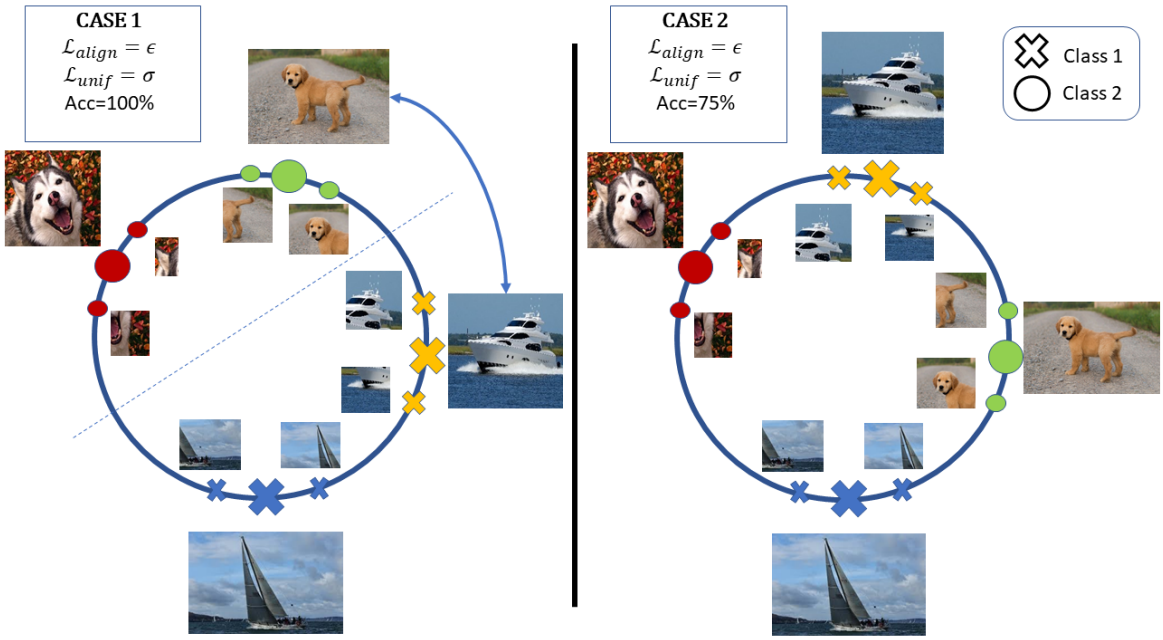


Figure 3.11: Alignment and Uniformity are necessary but not sufficient properties to produce a linearly separable latent space between classes. In both case 1 and 2, alignment and uniformity have the same values (ϵ and σ resp.) however only the left representation linearly separates boats from dogs with perfect accuracy. Why does CL leads to case 1 over case 2 ? We need additional assumptions on data augmentation and/or the family of encoders $\{f_\theta\}$ to answer.

Sadly, the answer is no in general since perfect alignment means that all positive samples are mapped to the same representation. Nevertheless, as empirically showed by Wang and Isola on vision datasets, both alignment and uniformity nicely correlate with downstream performance. It seems to be two necessary properties for having good representation. But is it sufficient ? Again, it is not (see Fig. 3.11). Intuitively, uniformity tries to repel all data samples from one another in the latent space to avoid big holes or clusters in some regions. It thus shapes the global latent space structure. It does not tell us anything about the *local* representation structure (as illustrated in Fig. 3.11 where we can swap any pair of image representation without changing $\mathcal{L}_{unif}$). On the other hand, $\mathcal{L}_{align}$ attracts positive pairs so it should be this term that avoids falling into case 2 (where the local neighborhood contains image representation from different semantic classes). By attracting the positive samples, $\mathcal{L}_{align}$ imposes that semantically close samples are also close in the representation space. Nonetheless, this hypothesis is not explicitly stated in the previous theoretical framework. We present our first assumptions and theoretical results in the next section to better understand the role of this alignment term on generalization performance.

3.3.2 Provable guarantees of contrastive learning with augmentation graph

Previous notations have their limitation since i) they hide the augmentation strategy inside $p(V_1)$ and $p(V_2)$ distributions and ii) they are highly focused on 2 views, limiting the analysis.

In what follows, we introduce additional notations to study CL framework with a focus on alignment and uniformity properties.

Introduction of Decoupled Uniformity loss

Setup. From N original samples $(x_i)_{i\in[1..N]} \in \mathcal{X} \overset{\text{i.i.d.}}{\sim} p(X)$, we transform them to generate semantically similar *positive samples* in the augmentation space $\mathcal{V}$ using an augmentation module $\mathcal{A}$ that induces a distribution $p_{\mathcal{A}}(V|X)$ (where V represent a view of X). Concretely, for each x_i, we can sample views of x_i using $v \sim p_{\mathcal{A}}(V|x_i)$ (e.g., by applying color jittering, flip or crop with a given probability, depending on $\mathcal{A}$). For consistency, we assume $p_{\mathcal{A}}(X) = p(X)$ so that probability distributions $p_{\mathcal{A}}(V|X)$ and $p(X)$ induce a marginal distribution $p_{\mathcal{A}}(V)$ over $\mathcal{V}$. Given an anchor x_i, all views $v \sim p_{\mathcal{A}}(V|x_j)$ from different samples $x_j \neq x_i$ are considered as *negatives*. From previous notations, we notably have the following connection:

$$p(v_1, v_2) = \mathbb{E}_{x\sim p(X)} \left(p_{\mathcal{A}}(v_1|x)p_{\mathcal{A}}(v_2|x)\right) \tag{3.26}$$

Linear evaluation. Once pre-trained, the encoder $f_\theta : \mathcal{V} \to \mathbb{S}^{d-1}$ is fixed and its representation $f_\theta(\mathcal{X})$ is evaluated[10] through linear evaluation on a classification task using a labeled dataset $\mathcal{D} = \{(x_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = [1..K]$, with K the number of classes. We note $\mathcal{F} = \{f_\theta\}$ the family of encoders. In practice, we train a linear classifier $g(x) = W f_\theta(x)$ (θ is fixed) that minimizes the multi-class classification error to perform linear evaluation.

Rethinking CL loss. The popular InfoNCE loss [211, 222], often used in CL, imposes 1) alignment between positives and 2) uniformity between the views of all instances $(x_i)_{i\in[1..N]}$ [297] – two properties that correlate well with downstream performance. However, by imposing uniformity between *all* views, we essentially try to both attract (alignment) and repel (uniformity) positive samples and therefore we cannot achieve a perfect alignment *and* uniformity, as noted in [297]. Moreover, InfoNCE has been originally designed for only two views (i.e., one couple of positive) and its extension to multiple views is not straightforward. Previous works have proposed a solution to either the first [276] or second [309] issue. Here, we propose a modified version of the uniformity loss $\mathcal{L}_{unif}$ (see previous section 3.3.1) that solves both issues since it: i) decouples positives from negatives, similarly to [309] and ii) is generalizable to multi-views as in [276]. We introduce the Decoupled Uniformity loss for $f \in \mathcal{F}$ as:

$$\mathcal{L}^d_{unif}(f) = \log \mathbb{E}_{(x,x')\sim p(X)p(X')} \exp\left(-||\mu_x - \mu_{x'}||^2\right) \tag{3.27}$$

where $\mu_x = \mathbb{E}_{v\sim p_{\mathcal{A}}(V|x)} f(v)$ is called a *centroid* of the views of x. This loss essentially repels distinct centroids μ_x through an average pairwise Gaussian potential. Interestingly, it implicitly optimizes alignment between positives through the maximization of $||\mu_x||$[11], so we do not need

[10] We assumed that $\mathcal{X} \subset \mathcal{V}$ which is true in practice since identity transformation is a possible augmentation.
[11] By Jensen's inequality $||\mu_x|| \leq \mathbb{E}_{v\sim p_{\mathcal{A}}(V|x)}||f(v)|| = 1$ with equality iff f is constant on $\operatorname{supp} p_{\mathcal{A}}(\cdot|x)$.

to explicitly add an alignment term. It can be shown (see Appendix B.5), that minimizing this loss brings to a representation space where the sum of similarities between views of the same sample is greater than the sum of similarities between views of different samples. From a physics point-of-view, we are trying to find the equilibrium state of $|\mathcal{X}|$ particles linked with a pairwise Gaussian potential energy. We will study its main properties hereafter and we will see that *prior* information can be added during the estimation step of these centroids.

Geometrical analysis of decoupled uniformity

Definition 3.3.1. *(Finite-samples estimator) For N variables $(x_i)_{i\in[1..N]} \overset{i.i.d.}{\sim} p(X)$, the (biased) estimator of $\mathcal{L}^d_{unif}(f)$ is: $\hat{\mathcal{L}}^d_{unif}(f) = \log \frac{1}{N(N-1)} \sum_{i\neq j} \exp(-||\mu_{x_i} - \mu_{x_j}||^2)$. It converges to $\mathcal{L}^d_{unif}(f)$ with rate $O\left(N^{-1/2}\right)$. Proof in Appendix B.8.*

Theorem 2. *(Optimality of Decoupled Uniformity) Given N points $(x_i)_{i\in[1..N]}$ such that $N \leq d+1$, any optimal encoder f^* minimizing $\hat{\mathcal{L}}^d_{unif}$ achieves a representation s.t.:*
1. (Perfect uniformity) All centroids $(\mu_{x_i})_{i\in[1..N]}$ make a regular simplex on the hyper-sphere $\mathbb{S}^{d-1}$
2. (Perfect alignment) f^* is perfectly aligned, i.e $\forall v_1, v_2 \sim p_{\mathcal{A}}(V|x_i), f^*(v_1) = f^*(v_2)$ for all $i \in [1..N]$.
Proof in Appendix B.8.

Theorem 2 gives a complete geometrical characterization when the batch size N set during training is not too large compared to the representation space dimension d. By removing the coupling between positives and negatives, we see that Decoupled Uniformity can realize both perfect alignment and uniformity, contrary to InfoNCE.
Remark. The assumption $N \leq d+1$ is crucial to have the existence of a regular simplex on the hypersphere $\mathbb{S}^{d-1}$. In practice, this condition is not always full-filled (e.g SimCLR [52] with $d = 128$ and $N = 4096$). Characterizing the optimal solution of $\mathcal{L}^d_{unif}$ for any $N > d+1$ is still an open problem [33] but theoretical guarantees can be obtained in the limit case $N \to \infty$ (see below).

Theorem 3. *(Asymptotical Optimality) When the number of samples is infinite $N \to \infty$, then for any perfectly aligned encoder $f \in \mathcal{F}$ that minimizes $\mathcal{L}^d_{unif}$, the centroids μ_x for $x \sim p(X)$ are uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$. Proof in Appendix B.8.*

Empirically, we observe that minimizers f of $\hat{\mathcal{L}}^d_{unif}$ remain well-aligned when $N > d+1$ on real-world vision datasets (see Fig. 3.12). Decoupled uniformity thus optimizes two properties that are nicely correlated with downstream classification performance [297]–that is alignment and uniformity between centroids. However, as we previously argued, optimizing these two properties is necessary but not sufficient to guarantee a good classification accuracy. In fact, the accuracy can be arbitrary bad even for perfectly aligned and uniform encoders (formal proof

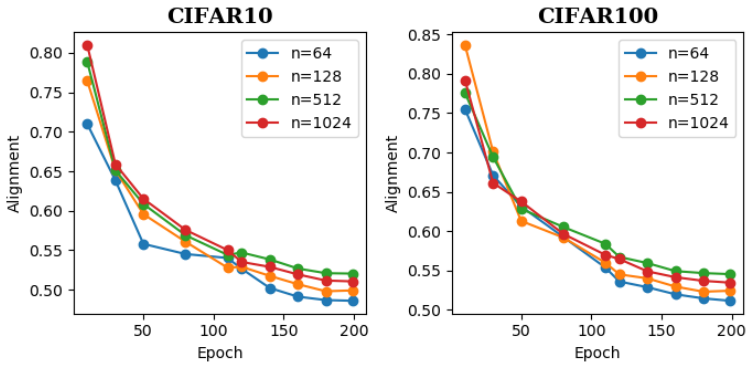


Figure 3.12: Decoupled Uniformity optimizes alignment, even in the regime when the batch size $N > d + 1$ (d=latent space dimension). Alignment metric $\mathcal{L}_{align}$ is computed on the validation set during optimization of Decoupled Uniformity loss with various batch sizes N and a fixed $d = 128$. We use 100 positive samples per image to compute $\mathcal{L}_{align}$ and SimCLR augmentations for module $\mathcal{A}$.

in [244] based on the same idea as depicted Fig. 3.11). Ultimately, it highly depends on the augmentation module $\mathcal{A}$, as we shall see.

Intuition. Most recent theories about CL [133, 300] make the hypothesis that samples from the same semantic class have overlapping augmented views to provide guarantees on the downstream task when optimizing InfoNCE or Spectral Contrastive loss [133]. This assumption, known as *intra-class connectivity hypothesis*, is very strong and only relies on the augmentation module $\mathcal{A}$. In particular, augmentations should not be "too weak", so that all intra-class samples are connected among them, and at the same time not "too strong", to prevent connections between inter-class samples and thus preserve the semantic information. Here, we prove that we can relax this hypothesis if we can provide a kernel (viewed as a similarity function between original samples $x \in \mathcal{X}$) that is "good enough" to relate intra-class samples not connected by the augmentations (see Fig. 3.13). In practice, we show that representation capacity of generative models can define such kernel.

We first recall the definition of the augmentation graph [300], and intra-class connectivity hypothesis before presenting our main theorems. For simplicity, we assume that the set of images $\mathcal{X}$ is finite (similarly to [133, 300]). Our bounds and theoretical guarantees will never depend on the cardinality $|\mathcal{X}|$.

Generalization guarantee under intra-class connectivity hypothesis

Definition 3.3.2. *(Augmentation graph [133, 300]) Given a set of original images $\mathcal{X}$, we define the augmentation graph $G_{\mathcal{A}}(V_e, E)$ for an augmentation module $\mathcal{A}$ through 1) a set of vertices $V_e = \mathcal{X}$ and 2) a set of edges E such that $(x, x') = e \in E$ if the two original images x, x' can be transformed into the same augmented image through $\mathcal{A}$, i.e* $\operatorname{supp} p_{\mathcal{A}}(\cdot|x) \cap \operatorname{supp} p_{\mathcal{A}}(\cdot|x') \neq \emptyset$.

Previous analysis in CL make the hypothesis that it exists an optimal (accessible) augmentation module $\mathcal{A}^*$ that fulfills:

Assumption 2. *(Intra-class connectivity [300]) For a given downstream classification task $\mathcal{D} = \{(x_i, y_i)\} \in \mathcal{X} \times \mathcal{Y}$ and any class $y \in \mathcal{Y}$, the augmentation subgraph, $G_y \subset G_{\mathcal{A}^*}$ containing images only from class y in $G_{\mathcal{A}^*}$, is connected.*

Under this hypothesis, Decoupled Uniformity loss can tightly bound the downstream supervised risk for a bigger class of encoders than prior work [300]. To show it, we define a measure of the risk on a downstream task $\mathcal{D}$. While previous analysis [11, 300] generally used the mean cross-entropy loss (as it has closer analytic form with InfoNCE), we use a supervised loss closer to decoupled uniformity with the same guarantees as the mean cross-entropy loss (see Appendix). Notably, the geometry of the representation space at optimum is the same as cross-entropy and SupCon [170] and we can theoretically achieve perfect linear classification.

Definition 3.3.3. *(Downstream supervised loss) For a given downstream task $\mathcal{D}$, we define the classification loss as: $\mathcal{L}_{sup}(f) = \log \mathbb{E}_{y,y' \sim p(Y)p(Y')} \exp(-||\mu_y - \mu_{y'}||^2)$, where $\mu_y = \mathbb{E}_{x \sim p(X|Y=y)} \mu_x$.*

Remark. This loss depends on centroids μ_x rather than $f(x)$. Empirically, it has been shown [102] that performing feature averaging gives better performance on the downstream task.

Definition 3.3.4. *(Weak-aligned encoder) An encoder $f \in \mathcal{F}$ is ϵ-weak ($\epsilon \geq 0$) aligned on $\mathcal{A}$ if:*

$$||f(x) - f(x')|| \leq \epsilon \qquad \forall x \in \mathcal{X}, \forall v_1, v_2 \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V|x)$$

Theorem 4. *(Guarantees with $\mathcal{A}^*$) Given an optimal augmentation module $\mathcal{A}^*$ that full-fills intra-class connectivity for a task $\mathcal{D}$, for any ϵ-weak aligned encoder $f \in \mathcal{F}$ we obtain:*

$$\mathcal{L}^d_{unif}(f) \leq \mathcal{L}_{sup}(f) \leq 8D\epsilon + \mathcal{L}^d_{unif}(f) \tag{3.28}$$

where D is the maximum diameter of all intra-class graphs G_y ($y \in \mathcal{Y}$). Proof in Appendix B.8.

In practice, the diameter D can be controlled by a small constant in some cases [300] (typically ≤ 4) but it remains specific to the dataset at hand. Furthermore, we observe in Fig. 3.12 that f realizes alignment with small error ϵ during optimization of $\mathcal{L}^d_{unif}(f)$ for augmentations close to the sweet spot $\mathcal{A}^*$ [276] on CIFAR-10 and CIFAR-100 (here $\mathcal{A}$ =SimCLR augmentations).

In the next section, we study the case when $\mathcal{A}^*$ is not accessible or very hard to find.

3.3.3 Reconnect the disconnected: extending the augmentation graph with kernel

Having access to optimal augmentations is a strong assumption and, for many real-world applications (e.g medical imaging [87]), it may not be accessible. If we have only weak augmentations (*e.g.*, $\operatorname{supp} p_{\mathcal{A}}(\cdot|x) \subsetneq \operatorname{supp} p_{\mathcal{A}^*}(\cdot|x)$ for any x), then some intra-class points might not

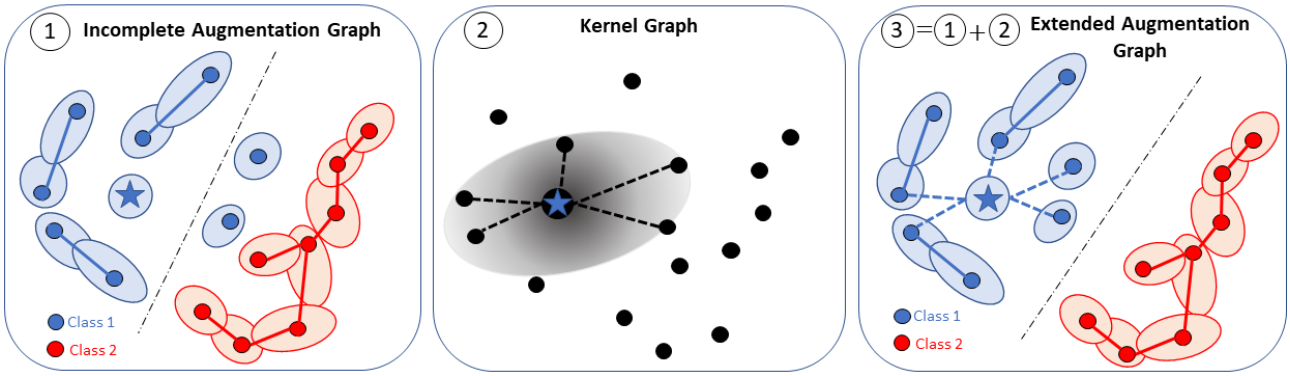


Figure 3.13: Illustration of the proposed method. Each point is an original image $x \in \mathcal{X}$. Two points are connected if they can be transformed into the same augmented image using a distribution of augmentations $p_{\mathcal{A}}$. Colors represent semantic (unknown) classes and light disks represent the support of augmentations for each sample x, $\operatorname{supp} p_{\mathcal{A}}(\cdot|x)$. From an incomplete augmentation graph (1) where intra-class samples are not connected (e.g. augmentations are insufficient or not adapted), we reconnect them using a kernel defined on prior information (either learnt with generative model, viewed as feature extractor, or given as auxiliary attributes). The extended augmentation graph (3) is the union between the (incomplete) augmentation graph (1) and the kernel graph (2). In (2), the gray disk indicates the set of points that are close to the anchor (blue star) in the kernel space.

be connected and we would need to reconnect them to ensure good downstream accuracy (see Theorem 10 in Appendix). Augmentations are intuitive and they have been hand-crafted for decades by using human perception (*e.g.*, a rotated chair remains a chair and a gray-scale dog is still a dog). However, we may know other *prior information* about objects that are difficult to transfer through invariance to augmentations (*e.g.*, chairs should have 4 legs). This prior information can be either given as image attributes (*e.g.*, age or sex of a person, color of a bird, etc.) or, in an unsupervised setting, directly learnt through a generative model (*e.g.*, GAN or VAE). Now, we ask: how can we integrate this information inside a contrastive framework to reconnect intra-class images that are actually disconnected in $G_{\mathcal{A}}$? We rely on conditional mean embedding theory and use a kernel defined on the prior representation/information. This allows us to estimate a better configuration of the centroids in the representation space, with respect to the downstream task, and, ultimately, provide theoretical guarantees on the classification risk.

Kernel Graph

Definition 3.3.5. *(RKHS on $\mathcal{X}$) We define the RKHS $(\mathcal{H}_{\mathcal{X}}, K_{\mathcal{X}})$ on $\mathcal{X}$ associated with a kernel $K_{\mathcal{X}}$.*

Example. If we work with large natural images, assuming that we know a prior $z(x)$ about our images (*e.g.,* internal representation of a generative model), then we can compute $K_{\mathcal{X}}$ using z as $K_{\mathcal{X}}(x, x') = \tilde{K}(z(x), z(x'))$ where $\tilde{K}$ is a standard kernel (*e.g.,* , Gaussian or Cosine).

To link kernel theory with the previous augmentation graph, we need to define a *kernel graph* that connects images with high similarity in the kernel space.

Definition 3.3.6. *(Kernel graph) Let $\epsilon > 0$. We define the ϵ-kernel graph $G^{\epsilon}_{K_{\mathcal{X}}}(V_e, E_{K_{\mathcal{X}}})$ for the kernel $K_{\mathcal{X}}$ on $\mathcal{X}$ through 1) a set of vertices $V_e = \mathcal{X}$ and 2) a set of edges $E_{K_{\mathcal{X}}}$ such that $e \in E_{K_{\mathcal{X}}}$ between $x, x' \in \mathcal{X}$ iff $\max(K_{\mathcal{X}}(x,x), K_{\mathcal{X}}(x',x')) - K_{\mathcal{X}}(x,x') \leq \epsilon$.*

The condition $\max(K_{\mathcal{X}}(x,x), K_{\mathcal{X}}(x',x')) - K_{\mathcal{X}}(x,x') \leq \epsilon$ implies that $d_{K_{\mathcal{X}}}(x,x') \leq 2\epsilon$ where $d_{K_{\mathcal{X}}}(x,x') = K_{\mathcal{X}}(x,x) + K_{\mathcal{X}}(x',x') - 2K_{\mathcal{X}}(x,x')$ is the kernel distance. For kernels with constant norm (*e.g.,* , the standard Gaussian, Cosine or Laplacian kernel), it is in fact an equivalence. Intuitively, it means that we connect two original points in the kernel graph if they have small distance in the kernel space. We give now our main assumption to derive a better estimator of the centroid μ_x in the insufficient augmentation regime.

Assumption 3. *(Extended intra-class connectivity) For a given task $\mathcal{D}$, the extended graph $\tilde{G} = G_{\mathcal{A}} \cup G^{\epsilon}_{K_{\mathcal{X}}} = (V, E \cup E_{K_{\mathcal{X}}})$ (union between augmentation graph and ϵ-kernel graph) is class-connected for all $y \in \mathcal{Y}$.*

This assumption is notably weaker than Assumption 2 w.r.t augmentation distribution $\mathcal{A}$. Here, we do not need to find the optimal distribution $\mathcal{A}^*$ as long as we have a kernel $K_{\mathcal{X}}$ such that disconnected points in the augmentation graph are connected in the ϵ-kernel graph. If $K_{\mathcal{X}}$ is not well adapted to the data-set (i.e it gives very low values for intra-class points), then ϵ needs to be large to re-connect these points and we will see that the classification error will be high. In practice, this means that we need to tune the hyper-parameter of the kernel (i.e., σ for a RBF kernel) so that all intra-class points are reconnected with a small ϵ.

Conditional Mean Embedding

Decoupled Uniformity loss includes no kernel in its original form. It only depends on centroids $\mu_x = \mathbb{E}_{v \sim p_{\mathcal{A}}(V|x)} f(v)$. Here, we show that another consistent estimator of these centroids can be defined, using the previous kernel $K_{\mathcal{X}}$. To show it, we **fix** an encoder $f \in \mathcal{F}$ and require the following technical assumption in order to apply conditional mean embedding theory [175, 263].

Assumption 4. *(Expressivity of $K_{\mathcal{X}}$) The (unique) RKHS $(\mathcal{H}_f, K_f)$ defined on $\mathcal{V}$ with kernel $K_f = \langle f(\cdot), f(\cdot) \rangle_{\mathbb{R}^d}$ fulfills $\forall g \in \mathcal{H}_f, \mathbb{E}_{v \sim p_{\mathcal{A}}(V|\cdot)} g(v) \in \mathcal{H}_{\mathcal{X}}$*

Theorem 5. *(Centroid estimation) Let $(v_i, x_i)_{i \in [1..N]} \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V, X)$. Assuming 4, a consistent estimator of the centroid is:*

$$\forall x \in \mathcal{X}, \hat{\mu}_x = \sum_{i=1}^{N} \alpha_i(x) f(v_i) \tag{3.29}$$

where $\alpha_i(x) = \sum_{j=1}^{n} [(K_N + N\lambda \mathbf{I}_N)^{-1}]_{ij} K_{\mathcal{X}}(x_j, x)$ and $K_N = [K_{\mathcal{X}}(x_i, x_j)]_{i,j \in [1..N]}$. It converges to μ_x with the ℓ_2 norm at a rate $O(N^{-1/4})$ for $\lambda = O(N^{-1/2})$. Proof in Appendix B.8.

Intuition. This theorem says that we can use representation of images close to an anchor x, according to our prior information, to accurately estimate μ_x. Consequently, if the prior

is "good enough" to connect intra-class images disconnected in the augmentation graph (i.e. fulfills Assumption 3), then this estimator allows us to tightly control the classification risk . From this theorem, we naturally derive the empirical Kernel Decoupled Uniformity loss using the previous estimator.

Definition 3.3.7. *(Empirical Kernel Decoupled Uniformity Loss) Let* $(v_i, x_i)_{i\in[1..N]} \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V, X)$. *Let* $\hat{\mu}_{x_j} = \sum_{i=1}^{N} \alpha_{i,j} f(v_i)$ *with* $\alpha_{i,j} = ((K_N + \lambda N \mathbf{I}_N)^{-1} K_N)_{ij}$, $\lambda = O(N^{-1/2})$ *a regularization constant and* $K_N = [K_{\mathcal{X}}(x_i, x_j)]_{i,j\in[1..N]}$. *We define the empirical kernel decoupled uniformity loss as:*

$$\hat{\mathcal{L}}^{d}_{unif}(f) \overset{\text{def}}{=} \log \frac{1}{N(N-1)} \sum_{i,j=1}^{N} \exp(-||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2) \tag{3.30}$$

Extension to multi-views. If we have L views $(v_i^{(l)})_{l\in[1..L]} \overset{\text{i.i.d.}}{\sim} p(V|x_i)$ for each x_i, we can easily extend the previous estimator with $\hat{\mu}_{x_i} = \frac{1}{L}\sum_{l=1}^{L} \hat{\mu}_{x_j}^{(l)}$ where $\hat{\mu}_{x_j}^{(l)} = \sum_{i=1}^{N} \alpha_{i,j} f(v_i^{(l)})$.

The computational cost added is roughly $O(N^3)$ (to compute the inverse matrix of size $N \times N$) but it remains negligible compared to the back-propagation time using classical stochastic gradient descent. Importantly, the gradients associated to $\alpha_{i,j}$ are not computed.

Generalization guarantees

We show here that $\hat{\mathcal{L}}^{d}_{unif}(f)$ can tightly bound the supervised classification risk for well-aligned encoders $f \in \mathcal{F}$.

Theorem 6. *We assume 3 and 4 hold for a reproducible kernel* $K_{\mathcal{X}}$ *and augmentation module* $\mathcal{A}$. *Let* $(v_i, x_i)_{i\in[1..N]} \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V, X)$. *For any* α*-weak aligned encoder* $f \in \mathcal{F}$*:*

$$\hat{\mathcal{L}}^{d}_{unif}(f) - O\left(N^{-1/4}\right) \leq \mathcal{L}_{sup}(f) \leq \hat{\mathcal{L}}^{d}_{unif}(f) + 4D(2\alpha + \beta_N(K_{\mathcal{X}})\epsilon) + O\left(N^{-1/4}\right) \tag{3.31}$$

where $\beta_N(K_{\mathcal{X}}) = (\frac{\lambda_{min}(K_N)}{\sqrt{N}} + \sqrt{N}\lambda)^{-1} = O(1)$ *for* $\lambda = O(N^{-1/2})$, $K_N = (K_{\mathcal{X}}(x_i, x_j))_{i,j\in[1..N]}$ *and* D *is the maximal diameter of all sub-graphs* $\tilde{G}_y \subset \tilde{G}$ *in the extended graph where* $y \in \mathcal{Y}$. *We noted* $\lambda_{min}(K_N) > 0$ *the minimal eigenvalue of* K_N.

Interpretation. Theorem 6 gives a tight bound on the classification loss $\mathcal{L}_{sup}(f)$ with few assumptions. In the special case $\epsilon = 0$ and $\mathcal{A} = \mathcal{A}^*$ (i.e the augmentation graph is class-connected, a stronger assumption than 3), we retrieve the standard bounds of Theorem 4. As before, we don't require perfect alignment for $f \in \mathcal{F}$ and we don't have class collision term (even if the extended augmentation graph may contain edges between inter-class samples), contrarily to [11]. Also, the estimation error doesn't depend on the number of views (which is low in practice))–as it was always the case in previous formulations [11, 133, 300] – but rather on the batch size N. Contrarily to CCLK [280], we don't condition our representation to weak attributes but rather we provide better estimation of the conditional mean embedding conditionally to the original image. Our loss remains in an unconditional contrastive framework driven by the augmentations $\mathcal{A}$ and the prior $K_{\mathcal{X}}$ on input images.

3.3.4 Experiments

Here, we study several problems where Kernel Decoupled Uniformity outperforms current contrastive models. In unsupervised learning, we show that we can leverage generative models representation to outperform current self-supervised models when the augmentations are insufficient to remove irrelevant signals from images. In a weakly supervised setting, we demonstrate the superiority of our unconditional formulation when noisy auxiliary attributes are available. Implementation details in Appendix B.7.

Generative models as prior - Evading feature suppression

Previous investigations [54] have shown that a few easy-to-learn irrelevant features not removed by augmentations can prevent the model from learning all semantic features inside images. We propose here a first solution to this issue.

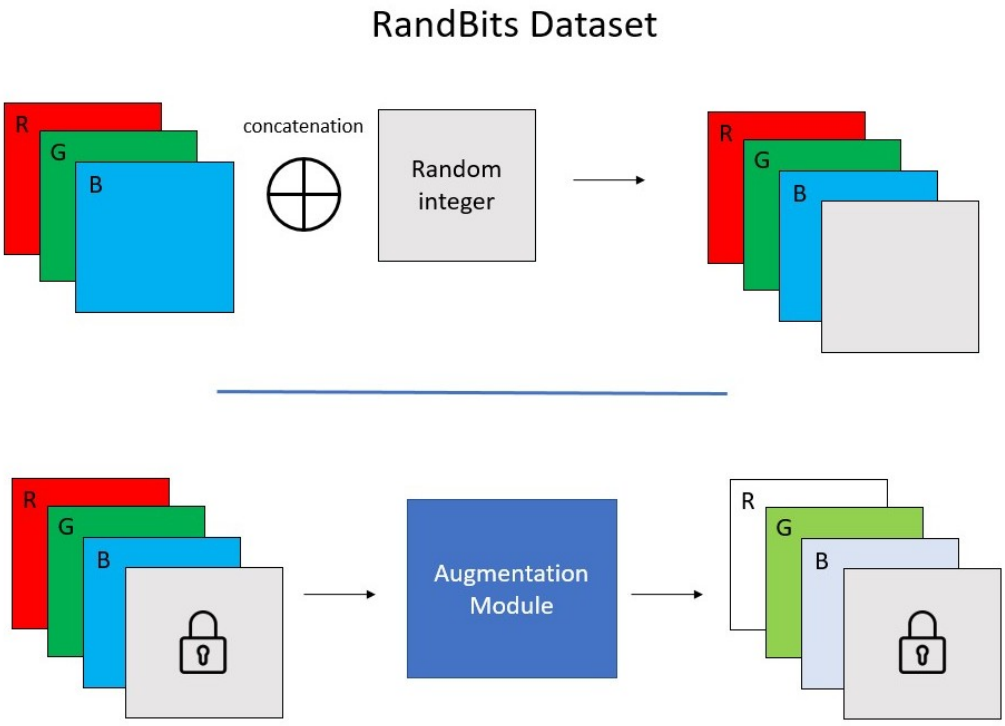


Figure 3.14: Illustration of RandBits dataset [54]. For each image, a random integer is added as an additional channel. The augmentation module $\mathcal{A}$ does not remove this noisy integer from images so it is shared between all views. In practice, the integer is randomly sampled between 0 and $2^k - 1$ with k the number of random bits. All CL models rely on this integer to perform their task, thus leading to poor representation. We provide a first solution using generative models as prior.

We build a RandBits dataset based on CIFAR-10 (see Fig. 3.14). For each image, we add a random integer sampled in $[0, 2^k - 1]$ where k is a controllable number of bits. To make it easy to learn, we take its binary representation and repeat it to define k channels that are added to the original RGB channels. Importantly, these channels will not be altered by augmentations, so they will be shared across views. We train a ResNet18 on this dataset with standard SimCLR

augmentations [52] and we make k vary. For Kernel Decoupled Uniformity loss, we use a β-VAE representation (ResNet18 backbone, $\beta = 1$) to define $K_{VAE}(x, x') = K(\mu(x), \mu(x'))$ where $\mu(\cdot)$ is the mean Gaussian distribution of x in the VAE latent space and K is a standard RBF kernel.

Loss	0 bits	5 bits	10 bits	20 bits
SimCLR [52]	79.4	68.74	13.67	10.07
BYOL [120]	80.14	19.98	10.33	10.00
β-VAE ($\beta = 1$)	41.37	43.32	42.94	43.1
β-VAE ($\beta = 2$)	42.28	43.89	43.11	42.19
β-VAE ($\beta = 4$)	42.5	42.5	42.5	39.87
Decoupled Unif (ours)	82.43	53.45	10.08	9.64
K_{VAE} Decoupled Unif (ours)	**82.74**$_{\pm 0.18}$	**68.75**$_{\pm 0.24}$	**68.42**$_{\pm 0.51}$	**68.58**$_{\pm 0.17}$

Table 3.5: Linear evaluation accuracy (in %) after training on RandBits-CIFAR10 with ResNet18 for 200 epochs. For VAE, we also use a ResNet18 backbone. Once trained, we use its representation to define the kernel K_{VAE} in Kernel Decoupled Uniformity loss.

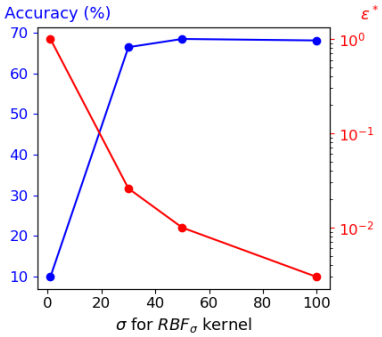


Figure 3.15: Empirical verification of our theory. The optimal ϵ^* to add 100 edges between intra-class images is correlated with the downstream accuracy, as suggested by Theorem 6. We use $k = 20$ bits and an RBF kernel.

Table 3.5 shows the linear evaluation accuracy computed on a fixed encoder trained with various contrastive (SimCLR, Decoupled Uniformity and Kernel Decoupled Uniformity) and non-contrastive (BYOL and β-VAE) methods. As noted previously [54], β-VAE is the only method insensitive to the number of added bits, but its representation quality remains low compared to other discriminative approaches. All contrastive approaches fail for $k \geq 10$ bits. This can be explained by noticing that, as the number of bits k increases, the number of edges between intra-class images in the augmentation graph $G_{\mathcal{A}}$ decreases. For k bits, on average $N/2^k$ images share the same random bits ($N = 50000$ is the dataset size). So only these images can be connected in $G_{\mathcal{A}}$. For $k = 20$ bits, < 1 image share the same bits which means that they are almost all disconnected, and it explains why standard contrastive approaches fail. Same trend is observed for non-contrastive approaches (*e.g.,* BYOL) with a degradation in performance even faster than SimCLR. Interestingly, encouraging a disentangled representation by imposing higher $\beta > 1$ in β-VAE does not help. Only our K_{VAE} Decoupled Uniformity loss obtains good scores, regardless of the number of bits.

Towards weaker augmentations

Color distortion (including color jittering and gray-scale) and crop are the two most important augmentations for SimCLR and other contrastive models to ensure good representation on ImageNet [52]. Whether they are best suited for other datasets (e.g medical imaging [86] or multi-objects images [54]) is still an open question. Here, we ask: can generative models remove the need for such strong augmentations ? We use standard benchmarking datasets (CIFAR-10,

CIFAR-100 and STL-10) and we study the case where augmentations are too weak to connect all intra-class points. We compare to baseline where all augmentations are used. We use a trained VAE to define K_{VAE} as before and a trained DCGAN [226] $K_{GAN}(x, x') \stackrel{\text{def}}{=} K(z(x), z(x'))$ where $z(\cdot)$ denotes the discriminator penultimate layer. In Table 3.6, we observe that our contrastive

Loss	CIFAR-10			CIFAR-100			STL-10		
	All	w/o Color	w/o Color +Crop	All	w/o Color	w/o Color +Crop	All	w/o Color	w/o Color +Crop
SimCLR [52]	79.4	62.56	34.07	49.50	38.27	15.28	76.99	59.01	39.56
BYOL [120]	80.14	64.86	45.88	51.57	35.61	22.48	77.62	65.36	11.28
Barlow Twins [314]	81.61	53.97	47.52	52.27	28.52	24.17	74.86	49.10	34.26
MoCo v3 [57]	**84.01**	67.71	42.12	**55.86**	36.95	22.11	**81.12**	64.25	38.38
VAE* [173]	41.37	41.37	41.37	14.34	14.34	14.34	42.17	42.17	42.17
DCGAN* [226]	66.71	66.71	<u>66.71</u>	26.17	26.17	26.17	70.06	<u>70.06</u>	**70.06**
Decoupled Unif (ours)	82.43	60.45	39.18	54.01	34.16	14.58	78.12	54.53	36.81
K_{VAE} Decoupled Unif (ours)	82.52	<u>72.92</u>	50.52	<u>54.66</u>	<u>45.59</u>	<u>28.24</u>	78.00	61.39	45.64
K_{GAN} Decoupled Unif (ours)	<u>83.01</u>	**77.16**	**69.19**	54.41	**50.07**	**35.98**	<u>78.50</u>	**71.44**	<u>68.11</u>

Table 3.6: When augmentation overlap hypothesis is not full-filled, generative models can provide a good kernel to connect intra-class points not connected by augmentations. * For VAE and DCGAN, we did not use augmentations during training since they model the true data distribution. Bold: best result; underlined: second best.

framework with DCGAN representation as prior is able to match the performance of SimCLR on CIFAR100 within 200 epochs by applying only crop augmentations and flip. Additionally, when removing almost all augmentations (crop and color distortion), we approach the performance of the prior representations of the generative models. This is expected by our theory since we have an augmentation graph that is almost disjoint for all points and thus we only rely on the prior to reconnect them.

ImageNet100. Current contrastive models do not match supervised performance on ImageNet. It means the augmentation graph is not entirely class-connected and there is still room for improvement. We show that BigBiGAN representation [84] provides a way to improve the performance of our contrastive model with standard SimCLR augmentations. First, to provide empirical evidence that decoupled uniformity loss (without kernel) is on par with current SOTA models, we optimize $\mathcal{L}^d_{unif}$ on 100-class subset of ImageNet (following [275]) in the multi-view setting. Then, we show that BigBiGAN encoder [84] pre-trained on ImageNet (without labels) can define a kernel $K_{GAN}(x, x') = K(z(x), z(x'))$ to improve contrastive-based model representation. K is an RBF kernel and $z(\cdot)$ is the BigBiGAN's encoder output.

Filling the gap for medical imaging

Data augmentations on natural images have been handcrafted over decades to achieve current performance on ImageNet. However, they might not be sufficient for medical datasets [87]. We study 1) bipolar disorder detection (BD), a challenging binary classification task, on brain MRI dataset BIOBD [148] and 2) chest radiography interpretation, a 5-class classification task on

Model	Epochs	ImageNet100
SimCLR [52] (repro)	400	66.52
BYOL [120] (repro)	400	72.26
CMC [275]	400	73.58
DCL [61]	400	74.6
AlignUnif [297]	240	74.6
Decoupled Unif (4 views)	400	74.70
BigBiGAN [84]	-	72.0
K_{GAN} Decoupled Unif (4 views)	400	**76.60**
Supervised	100	$82.1_{\pm 0.59}$

Table 3.7: Linear evaluation accuracy(%) with our model pre-trained on ImageNet100 using BigBiGAN representation trained on ImageNet as prior information for Decoupled Uniformity. We use ResNet50 trained on 400 epochs. Gray: ImageNet pre-training (w/o labels).

Model	BD vs HC
SimCLR [52]	$60.46_{\pm 1.23}$
BYOL [120]	$58.81_{\pm 0.91}$
MoCo v2 [136]	$59.27_{\pm 1.50}$
Model Genesis [320]	$59.94_{\pm 0.81}$
VAE [173]	$52.86_{\pm 1.24}$
K_{VAE} Decoupled Unif (ours)	$\mathbf{62.19}_{\pm 1.58}$
Supervised	$67.42_{\pm 0.31}$

Table 3.8: Linear evaluation AUC(%) for discriminating bipolar disorder vs controls using brain MRI and DenseNet121 model. All models are pre-trained on BHB-10K, a large dataset of brain scans from healthy controls. Standard deviation is reported with a 5-fold leave-site-out CV scheme to avoid possible bias on acquisition site.

Loss	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
SimCLR	82.42	77.62	90.52	89.08	86.83
BYOL	83.04	81.54	90.98	90.18	85.99
MoCo-CXR [264]	75.8	73.7	77.1	86.7	85.0
GLoRIA [150]	86.70	**86.39**	90.41	90.58	91.82
CCLK [280]	86.31	83.67	92.45	91.59	91.23
K_{GloRIA} Decoupled Unif (ours)	**86.92**	85.88	**93.03**	**92.39**	**91.93**
Supervised [35]	81.6	79.7	90.5	86.8	89.9

Table 3.9: AUC scores(%) under linear evaluation for discriminating 5 pathologies on CheXpert images. ResNet18 backbone is trained for 400 epochs (batch size $N = 1024$) without labels on official CheXpert training set and results are reported on validation set.

CheXpert [154]. BIOBD contains 356 healthy controls (HC) and 306 patients with BD. We use BHB-10K as a large pre-training dataset containing 10k 3D images of healthy subjects (as in section 3.2.3). For CheXpert, we use Gloria [150] representation, a multi-modal approach trained with (medical report, image) pairs to extract 2048-d features as weak annotations. We show that our approach improves contrastive model in both unsupervised (BD) and weakly supervised (CheXpert) setting for medical imaging.

Weakly supervised learning on natural images

Now we assume to have access to image attributes that correlate well with true semantic labels (e.g birds color or size for birds classification). We use three datasets: CUB-200-2011 [303], ImageNet100 [275] and UTZappos [312], following [280]. CUB-200-2011 contains 11788 images of 200 bird species with 312 binary attributes available (encoding size, wing shape, color, etc.). UTZappos contains 50025 images of shoes from several brands sub-categorized into 21 groups that we use as downstream classification labels. It comes with 7 attributes. Finally, for ImageNet100 we follow [280] and use the pre-trained CLIP [227] model (trained on pairs (text,

Loss	CUB	ImageNet100	UT-Zappos
SimCLR	17.48	65.30	84.08
BYOL	16.82	72.20	85.48
CosKernel CCLK [280]	15.61	74.34	83.23
RBFKernel CCLK [280]	30.49	77.24	84.65
CosKernel Decoupled Unif	27.77	**78.8**	**85.56**
RBFKernel Decoupled Unif	**32.87**	76.34	84.78

Table 3.10: If images attributes are accessible (e.g birds color or size for CUB200), they can be leveraged as prior in our framework to improve the representation.

image)) to extract 512-d features considered as prior information. We compare our method with CCLK, a conditional contrastive model that defines positive samples only according to the conditioning attributes.

Analysis of temperature and batch size on Decoupled Uniformity loss

InfoNCE is known to be sensitive to batch size and temperature to provide SOTA results. In our theoretical framework, we assumed that $f(x) \in \mathbb{S}^{d-1}$ but we can easily extend it to $f(x) \in \sqrt{t}\mathbb{S}^{d-1}$ where $t > 0$ is a temperature hyper-parameter. It defines the radius of the hypersphere and the corresponding loss function is $\mathcal{L}^d_{unif}(f) = \mathbb{E}_{(x,x')\sim p(X)p(X')} \exp\left(-t||\mu_x - \mu_{x'}||^2\right)$. In Table 3.11 and 3.11, we show that Decoupled Uniformity does not require large batch size (as it is the case for SimCLR with InfoNCE) and it produces good representations for $t \in [1, 5]$.

Datasets	$t = 0.1$	$t = 0.5$	$t = 1$	$t = 2$	$t = 5$	$t = 10$
CIFAR10	73.91	83.01	84.72	85.82	83.05	74.82
CIFAR100	39.16	51.33	55.91	58.89	56.70	48.29

Table 3.11: Linear evaluation accuracy (%) after training for 400 epochs with batch size $N = 256$ and varying temperature t in Decoupled Uniformity loss with SimCLR augmentations. $t = 2$ gives overall the best results, similarly to the uniformity loss in [297].

Datasets	Loss	$n = 128$	$n = 512$	$n = 1024$	$n = 2048$
CIFAR10	SimCLR	78.89	79.40	80.02	80.06
	Decoupled Unif	82.67	82.12	82.74	82.33
CIFAR100	SimCLR	49.53	53.46	54.45	55.32
	Decoupled Unif	54.61	54.12	55.56	55.20

Table 3.12: Linear evaluation accuracy (%) after training for 200 epochs with a batch size N, ResNet18 backbone and latent dimension $d = 128$. Decoupled Uniformity is less sensitive to batch size than SimCLR thanks to its decoupling between positives and negative samples.

3.3.5 Conclusion

This work was devoted to novel theoretical developments for contrastive learning (CL) leading to new generalization guarantees. In particular, we showed how prior information (e.g. given

by generative models) can define a prior structure in the representation space that can be ultimately leveraged to improve the final representation of images using DNN. We have drawn connections between kernel theory and CL to build our theoretical framework. As opposed to previous section 3.2, we did not rely on conditional independence hypothesis, but rather on the (weaker) intra-class connectivity hypothesis in the extended augmentation graph to derive tight bounds on downstream classification task. In practice, we show that generative models provide a good prior when augmentations are too weak or insufficient to remove easy-to-learn noisy features. We show applications to medical imaging in a fully unsupervised setup but also in the weakly supervised setting on natural images. We hope that CL will benefit from the future progress in generative modelling with our theoretical framework and it will widen its field of application to challenging tasks, such as computer aided-diagnosis.

This study is also an extension of our previous analysis where we only studied CL through Information Theory (IT) with a weakly supervised signal. We argued that IT does not provide a satisfying theoretical framework to study CL and we have based our analysis on metric learning instead, using concepts of *alignment* and *uniformity* for CL. Future work will consist in comparing the previous y-Aware InfoNCE estimator with Decoupled Uniformity loss and to analyze its theoretical property using the tools developed in this section, namely augmentation and kernel graph along with conditional mean embedding theory. Finally, our theory provides guarantees only for in-domain images, *i.e.,* images in pre-training and downstream tasks come from the same source domain. However, our main paradigm described Fig. 3.1 assumes that images on downstream tasks also come from out-domain, *i.e.,* from patients with brain pathology as opposed to healthy controls. Consequently, an important future direction is to study *linear transferability* (a concept proposed by HaoChen [134]) of an encoder pre-trained only on one domain (healthy controls) and whose representation is transferred to several other domains (e.g. psychiatric disorders).

Chapter 4

OpenBHB challenge for supervised representation learning and debiasing in neuroimaging

Contents

This work has been presented in:

- **OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing**
  B. Dufumier, A.Grigis, J. Victor, C. Ambroise, V. Frouin, E. Duchesnay
  NeuroImage, 2022
- **Supervised Contrastive Learning for Debiasing**
  C. A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, P. Gori
  ICLR 2023
- **Contrastive learning for regression in multi-site brain age prediction**, C. A. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, Pietro Gori
  ISBI 2023

With the growing emergence of new large-scale multi-site resource for neuroimaging (e.g. UKBioBank [37], HCP [286], etc.), we anticipate the emergence of deep models for supervised representation learning. However, as we saw in Chapter 2, these imaging data are often collected with different scanners and acquisition protocols, reflecting the inevitable constraints and objective of each neuroimaging study to answer broad questions in neuroscience (e.g. human brain development with HCP [286], aging with UKBioBank [37], biomarker discovery for ASD with ABIDE [79]). These discrepancies between studies highly influence image quality and induce a serious bias in machine learning (ML) models, a phenomena well described in Chapter 2. As D. Bzdok hypothesized [38]: *Across-site heterogeneity may explain why, counter-intuitively, predictive model performance have been repeatedly reported to decrease as the available neuroscience data increases [306].*

As an illustrative example, let us consider two cohorts, C1 and C2, acquired on two different scanners. We assume that C1 only contains males and C2 only females. Furthermore, we assume the two scanners have different permanent magnetic field (e.g. 1.5T and 3T) thus leading to different spatial resolutions. An ML algorithm trained to predict sex from $\{C1, C2\}$ can very well over-fit on spatial resolution quality instead of a neuroanatomical pattern to achieve perfect accuracy. If the train-test splits are stratified according to sex and scanner, this algorithms would even achieve good accuracy on test. We see that biased representation arises from a high correlation between the target to predict (sex in previous example) and the confounding variable (a.k.a bias which is the scanner in previous example). We argue that such bias limits the transfer capacity of pre-trained models and it can even lead to false discovery, especially for small sample size studies [196].

This chapter is devoted to tackle this issue. We first present a new large-scale brain sMRI resource-OpenBHB-publicly available, along with a machine learning challenge focused on supervised representation learning for brain age prediction with site-effect removal, viewed as a debiasing task. Accurately estimating biological age from brain imaging is an on-going challenge

which may provide important insights for biomarkers discovery and personalized medicine, as we shall discuss. OpenBHB is quite unique for its size (including $N > 5$k subjects) and its heterogeneity (71 acquisition centers spread worldwide over 3 continents- Asia, North America and Europe). It is focused on the healthy population and it is lifespan with standardized pre-processing pipelines for both surface-based and volume-based MRI analysis. We first study OpenBHB properties before presenting the challenge currently available on RAMP platform. This challenge introduces key metrics derived from submitted models representation (in particular through linear probing [5])), assessing their bias on acquisition site and their cross-site generalization performance for brain age prediction. Finally, we present first experiments from SOTA DNN models trained on several MRI modalities (whole-brain volume-based and surface-based measurements including gray matter volume, cortical thickness, surface area, local curvature etc.) and we evaluate SOTA harmonization model, namely ComBat [101].

4.1 Introduction

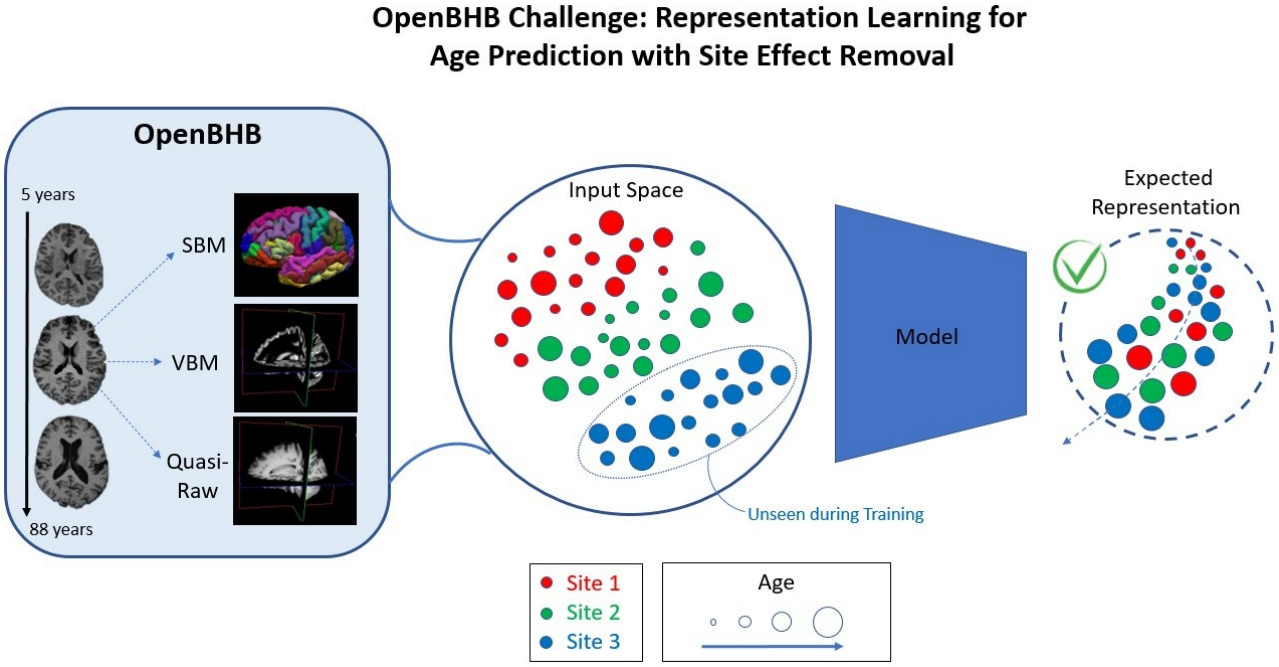


Figure 4.1: Illustration of the OpenBHB dataset along with the proposed challenge. OpenBHB is a large-scale ($N > 5$K subjects), international (covers Europe, North America, and China), lifespan (5-88 years old) brain MRI dataset including images preprocessed with three pipelines (quasi-raw, VBM with CAT12, and SBM with FreeSurfer). It is openly accessible on IEEE Dataport. It comes with a new challenge on representation learning for brain age prediction with site debiasing. Challenge information and dataset accessibility procedure are described on our website.

Brain aging implies several complex processes (e.g., cortical thinning or synaptic pruning) that vary drastically across individuals [213]. In particular, this maturation affects several functional and structural networks involved in cognition (e.g., working memory), motor func-

tions, or emotion (e.g., Default Mode Network). It has been shown [25, 65, 67, 164, 218] that machine learning (ML) models can learn from neuroimaging data to accurately estimate chronological age from the healthy population, taking into account the general variability (both environmental and genetic [164]) to build strong predictors of brain development. The Brain Age Gap (BAG, defined as the absolute difference between chronological and predicted age) has been used as a proxy measure to detect both neurodegenerative and neurodevelopmental disorders [110, 167, 178, 247] (e.g., Alzheimer or schizophrenia). It has also been described as a predictor of mortality [69] and other brain disorders (such as major depressive disorder [131], bipolar disorder [167] or traumatic brain injuries [66]).

Nevertheless, there are currently several shortcomings in the neuroimaging literature that heavily limit the clinical impact of such algorithms. First, the lack of public benchmarks necessarily limits the comparison and reproducibility of competing works on brain age prediction. Recent studies [137, 138] show that the choice of age range, number of samples, and pre-processing strategies - e.g. Region-Of-Interest (ROI), Voxel-Based Morphometry (VBM) or Surface-Based Morphometry (SBM) - are drastically different across studies, making the comparison difficult. In this regard, the Predictive Analytic Challenge [96] (PAC) in 2019 catalyzed the development of new ML algorithms and Deep Learning (DL) networks specially engineered for relatively large-scale ($N = 2636$) brain MRI data. Nevertheless, the development of DL architectures for neuroimaging data is still lagging behind the ones developed for natural images (e.g., there is still no consensus whether DL models are more efficient than simple regularized linear models [2, 249] on phenotype prediction even if more and more evidence is accumulating for the former [137, 138, 218]).

Second, most large emerging datasets are multi-sites (e.g., ABIDE, ABCD, ADNI, ENIGMA, SCHIZCONNECT), partly because of the high acquisition cost per patient in each study. Several recent works [112, 293] have shown that ML models are heavily biased by the acquisition site, and they generalize poorly to MRI images coming from never-seen sites. This issue can be attributed to the difference between scanners manufacturers, specifications, settings, and hardware.

This is an important limitation for applying these models to neuroimaging data, especially for personalized medicine in psychiatry. In this context, harmonization methods [101, 108, 171] have emerged to remove this undesired variability from the data. However, such harmonization models estimate their parameters on the entire dataset, or at least, on a great portion of it containing all sites. It is also a limitation in the context of personalized medicine, where MRI data coming from new hospitals would mean to re-train the whole model before making a new prediction. Besides, these methods are also sensitive to the number of samples per site as some statistics (mean or variance) are estimated for each site separately. Other recent approaches [24, 78, 193, 194, 235] are integrating DL to perform image-to-image translation (e.g style transfer) in order to bring all images in a common debiased space. Validating such approaches

is often difficult and it either relies on travelling patients (scanned at multiple sites), which is very costly, or on statistical analysis on the generated images (e.g using Fréchet Inception Distance [194]) or directly by demonstrating that biological variables are well preserved (e.g age or sex [24, 235]). Other line of work [82] directly tries to remove site information via adversarial attack while training an encoder to predict the biological variable of interest (e.g age or sex). In that case, the validation procedure simply consists in evaluating the encoder's capacity to retain biological and site information. All these approaches use different validation procedure and they are hardly comparable to one another (as they generally do not even use the same datasets and modalities).

As a result, we propose the OpenBHB Challenge on brain age prediction with site-effect removal. This challenge is based on the large-scale ($N > 5000$) multi-site brain MRI dataset OpenBHB that contains both minimally preprocessed data along with VBM and SBM measures derived from raw T1w MRI. All images in OpenBHB have passed a semi-automatic visual quality check, and the data are publicly available on the online IEEE Dataport platform. The challenge consists in learning a representation of the data such that i) brain age variability is preserved and ii) site-related information is removed. The submitted models should output a vector representing input data such that brain age can be easily predicted (i.e. through linear evaluation) and acquisition site signal is absent (i.e. random chance for predicting site with linear evaluation). Thus, this challenge is closely related to several hot topics in ML/DL, such as representation learning driven by a supervised signal [87, 170], debiasing and trustworthy AI [17, 21, 41, 62, 273]. To evaluate the submitted models, we propose a novel metric computed on two test sets: an *internal test* that contains images from the same sites as training and and *external test* including images from distinct sites. We hope this challenge will facilitate the benchmarking of ML and DL models for both brain age prediction and site-effect removal through a representation learning approach.

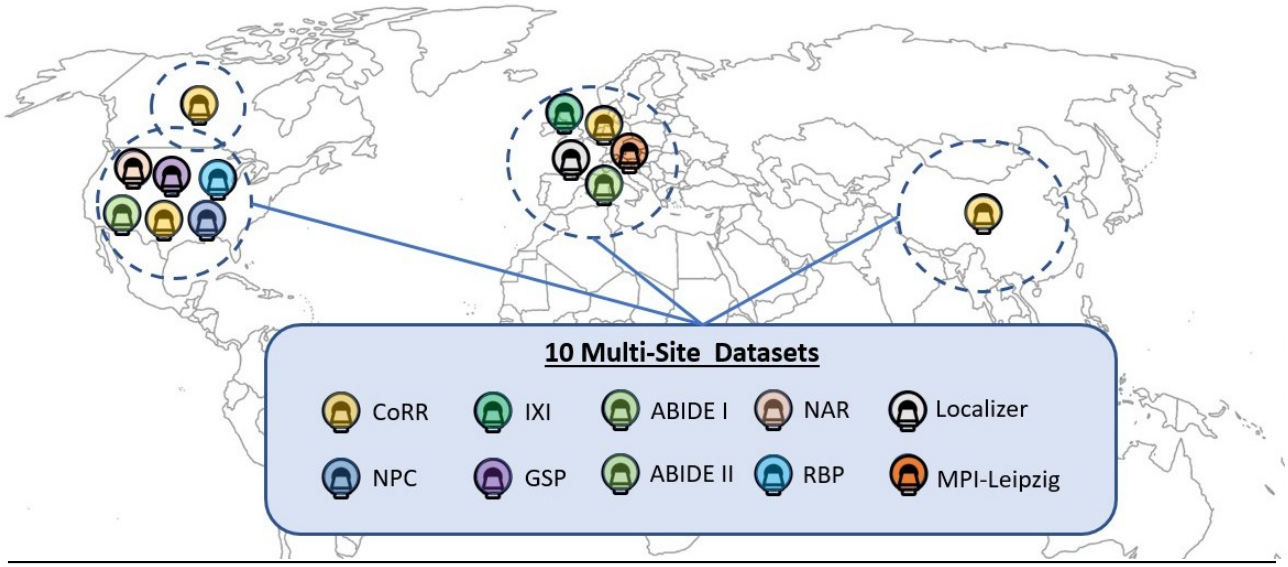
We plan to extend OpenBHB with additional subjects, longitudinal data and other modalities (e.g., resting-state functional MRI and diffusion MRI) that bring complementary structural and functional information to the current T1w images.

In summary, in this work our main contributions are:

- OpenBHB, a new large-scale ($N > 5000$) brain MRI dataset publicly available that includes:
  - preprocessed quasi-raw, VBM and SBM T1w data;
  - a visual quality check;
  - a training, validation and test splits used for the OpenBHB challenge;
- a new challenge for brain age prediction with site-effect removal.
- a leader-board for the comparison of submitted models with a new metric.
- an online platform to submit the trained models

4.2 OpenBHB dataset

4.2.1 Public datasets aggregated in OpenBHB



Study	# Subjects	# Scans	Age	Sex (%M)	# Sites	# Settings	Modalities (others available)
ABIDE I	527	527	17.2 ± 7.7	82.0	20	1	T1w, (rfMRI)
ABIDE II	555	555	14.9 ± 9.5	69.5	17 (1 out)	1	T1w (rfMRI)
CoRR	1368	1368	25.9 ± 15.9	49.4	26 (8 out)	1	T1w, (rfMRI)
GSP	1570	1570	21.5 ± 2.9	42.4	5	1	T1w, (rfMRI)
IXI	558	558	48.7 ± 16.4	44.4	1	3	T1w, (T2w, DWI, PD)
Localizer	76	76	24.5 ± 6.6	42.7	2	1	T1w, (fMRI)
MPI-Leipzig	282	282	35.6 ± 17.8	60.3	1	2	T1w, (rfMRI)
NAR	289	289	22.1 ± 4.9	41.9	3	1	T1w, (T2w, fMRI)
NPC	64	64	26 ± 4.2	55	1	1	T1w, (T2w, DWI, rfMRI)
RBP	41	41	23.1 ± 5.0	48.8	1	1	T1w, (T2w, DWI, rfMRI)
Total	5330	5330	25.3 ± 15	52.1	71 (9 out)	13	T1w

Table 4.1: OpenBHB demographic information. Acquisition settings include mainly the magnetic field strength and acquisition protocol used for MRI acquisition (see Sec. 4.2.1 for more details). Only images with available acquisition settings are included in the OpenBHB challenge (the number of sites excluded for the challenge are indicated in parentheses). Six sites are shared between ABIDE 1 and 2, and only healthy subjects are considered in the current release.

OpenBHB aggregates 10 publicly available datasets, namely IXI[1], ABIDE 1 [79], ABIDE 2 [80], CoRR [324], GSP [36], LOCALIZER [212], MPI-Leipzig [16], NAR [206], NPC [269], and RBP [100, 192]. Currently, OpenBHB is focused only on Healthy Controls (HC) since the main challenge consists in modeling the (normal) brain development by building a robust brain age predictor. As a result, we only included HC from ABIDE 1 and 2, and we left out the subjects with Autism Spectrum Disorders in the current release. OpenBHB contains $N = 5330$ 3D T1 brain MRI scans from HC acquired on 71 different acquisition sites with eventually multiple acquisition protocols per site (see Tab. 4.1). As highlighted in the map accompanying Tab. 4.1, the subjects included in OpenBHB come from European-American, European, and Asian genetic backgrounds, promoting more diversity in OpenBHB. To manage

[1] https://brain-development.org/ixi-dataset

redundant images, one session per participant has been retained along with its best-associated run, selected according to image quality. We also provide the participants phenotype as well as site and scanner information associated with each image, which essentially includes age, sex, acquisition site, diagnosis (in our case only HC), MRI scanner magnetic field, and MRI scanner settings identifier (a combination of multiple information composed of a subset of the repetition time, echo time, sequence name, flip angle, and acquisition coil). Some widespread confounds are also proposed, such as the Total Intracranial Volume (TIV), the CerebroSpinal Fluid Volume (CSFV), the Gray Matter Volume (GMV), and the White Matter Volume (WMV).

The overall age and sex distributions for OpenBHB are plotted Fig. 4.3. It should be noticed that sex distribution is globally well balanced for all age bins. Age distribution contains 2 main modes centered around 10 years old (during synaptic pruning) and 25 years old with a long tail above 40 until 88 years (and fewer samples in this range).

Additionally, we performed appearance analysis from VBM data, as it preserves cortical and sub-cortical information from raw images (as opposed to SBM) and ROI measures can be derived easily to reduce data dimensionnality. Specifically, we represented the t-SNE visualization of ROI extracted from VBM data in Fig. 4.2 per study. This plot clearly suggests that age and site effect are driving the representation, thus justifying the objective of the OpenBHB challenge. Datasets with *both* a large number of sites and a large age range cover wider regions in t-SNE space than others (e.g CoRR covers almost all regions while GSP and IXI cover only middle and upper regions; even if they are lifespan, they only include 6 sites together vs 18 sites for CoRR). This is even more obvious with MPI-Leipzig that covers mostly small left and upper regions while it is also lifespan.

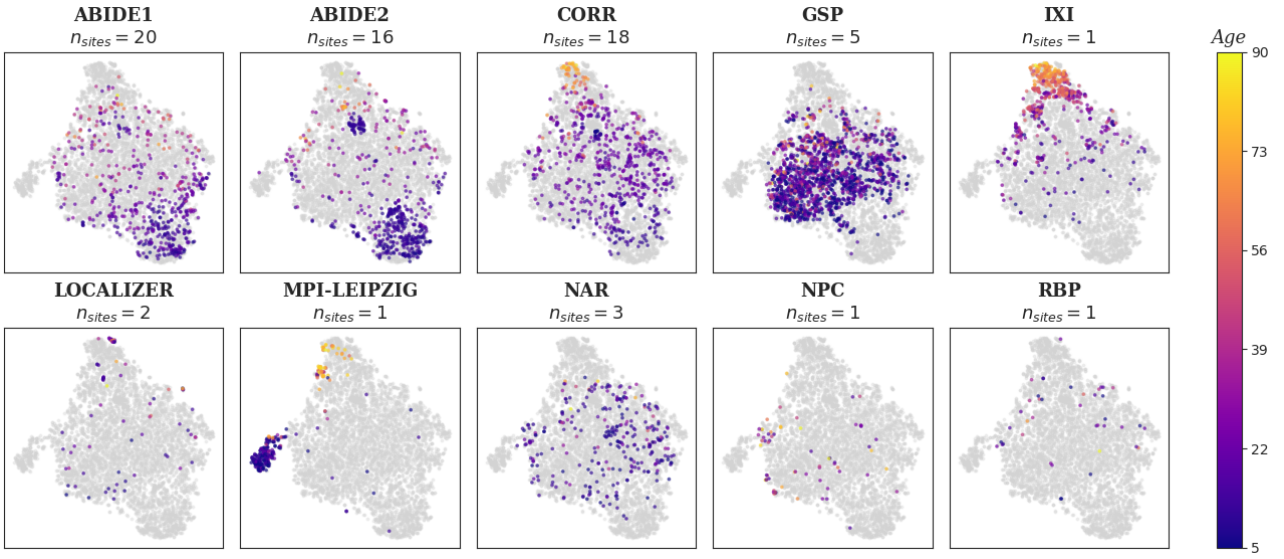
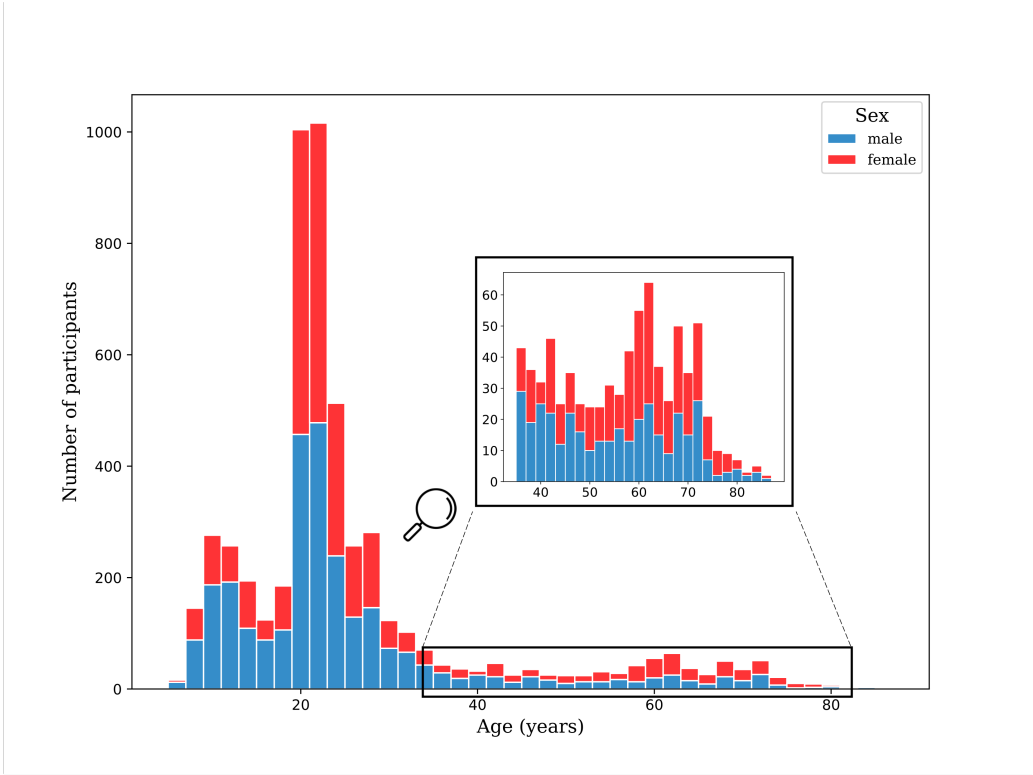
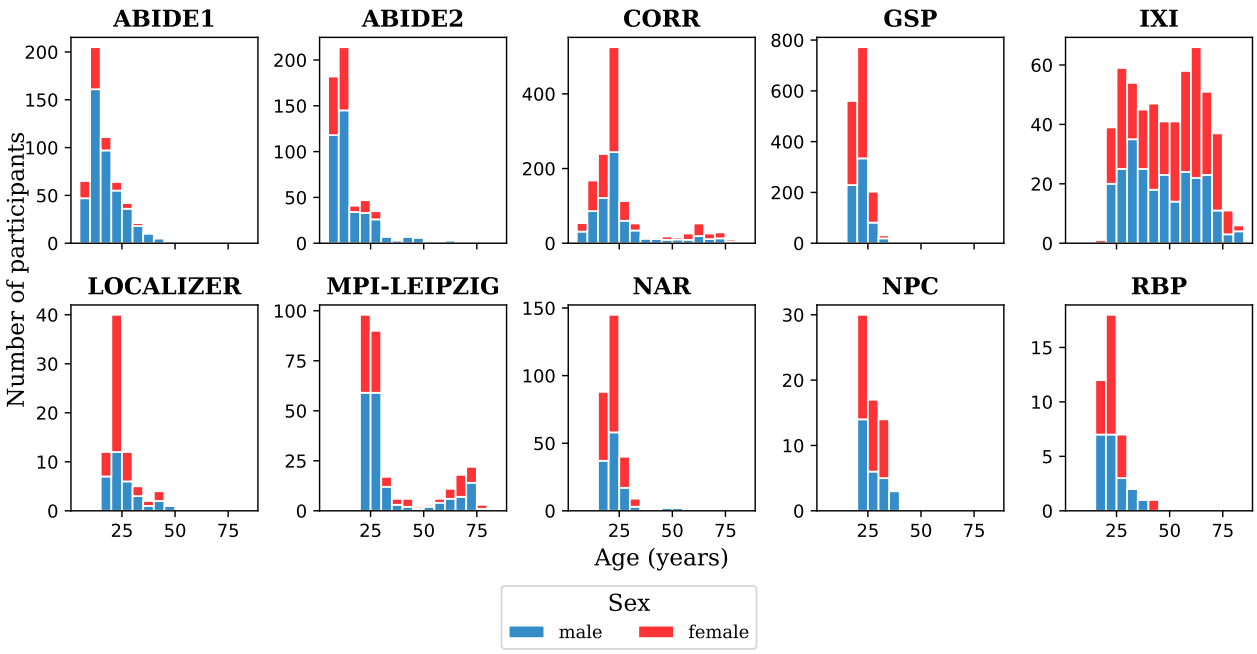


Figure 4.2: t-SNE representation of VBM ROI normalized by TIV for each sample in OpenBHB Challenge. Age and sites dominate the representation, and studies with multiple sites have a broader variety of images (in terms of surface covered in the t-SNE space). For instance, MPI-Leipzig covers mostly a single region on the left while NAR is more varied. CoRR is the most varied study as it covers almost all representation space.



(a) Overall age and sex distributions.



(b) Age and sex distributions by study

Figure 4.3: Demographic description of OpenBHB (a) overall and (b) by study. Age histograms show a peak distribution for young adults (20-30 years old) with a long tail distribution for older adults (60-80 years old). While age disparities are observed between studies, the age remains a poor site predictor (see Table 4.3). All data-sets are well-balanced between males and females.

4.2.2 Preprocessing and derived anatomical features

All data are preprocessed uniformly with container technologies comprising quasi-raw, CAT12 VBM, and FreeSurfer (see Fig. 4.2.2), which allows us to control the different software versions over time. The project hosting the codes is freely accessible at https://brainprep.readthedocs.io. We conducted a semi-automatic quality control (QC) guided with quality metrics leading to a selection of images that meet the quality criteria for all three pre-processing pipelines (see Fig. 4.5 and the detailed QC per pre-processing below).

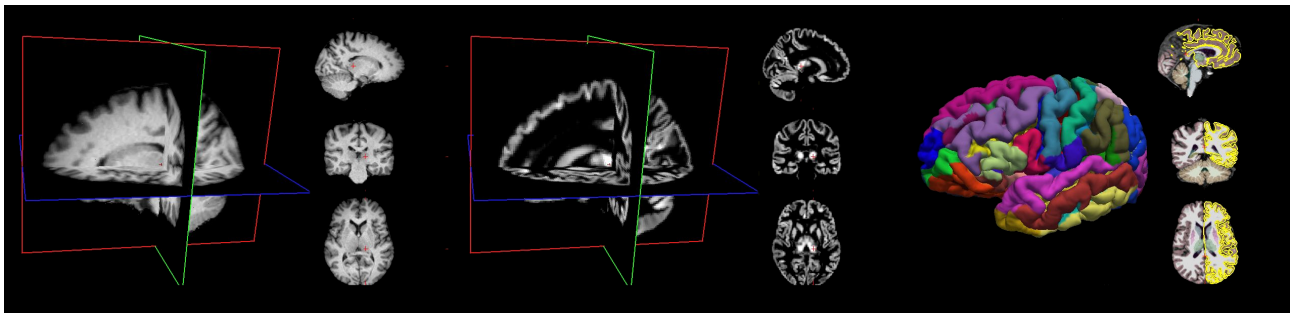


Figure 4.4: Illustration of the OpenBHB available preprocessed data: quasi-raw, Gray Matter (GM) CAT12 VBM, and FreeSurfer "recon-all" from left to right.

Quasi-raw

Steps: Minimally preprocessed data were generated using ANTS[13] bias field correction, FSL FLIRT[158] with 9 degrees of freedom (no shearing) followed by affine registration to the $1mm^3$ MNI template, and the application of a brain mask to remove non-brain tissues in the final images.

Quality control: First, we computed the correlation between each image and the the mean of every other images in order to sort them by increasing correlation score. Then, images were manually inspected in-house following this sorting and a first threshold was set to remove the first k images. Additionally, we used the average correlation (using Fisher's z transform) between registered images as a metric of quality and we retained only images at a threshold higher than 0.5 (see Fig. 4.5).

CAT12 VBM

Steps: Voxel-Based Morphometry (VBM) was performed with CAT12[109] (http://www.neuro.uni-jena.de/cat). The analysis stream includes non-linear spatial registration to the $1.5mm^3$ MNI template, Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF) tissues segmentation, bias correction of intensity non-uniformities, and segmentations modulation by scaling with the amount of volume changes due to spatial registration. VBM is most often applied to investigate the GM. The sensitivity of VBM in the WM is low, and

usually, diffusion-weighted imaging is preferred for that purpose. For this reason, only the modulated GM images are shared. Moreover, CAT12 computes GM volumes averaged on the Neuromorphometrics atlas that includes 284 brain cortical and sub-cortical ROI.

Quality control: We performed the same in-house QC visual analysis as for quasi-raw images (see section 4.2.2). Additionally, we also monitored the Noise Contrast Ratio (NCR) and Image Quality Rating (IQR) as two metrics of quality and we retained only images at a threshold below 4 (see Fig. 4.5).

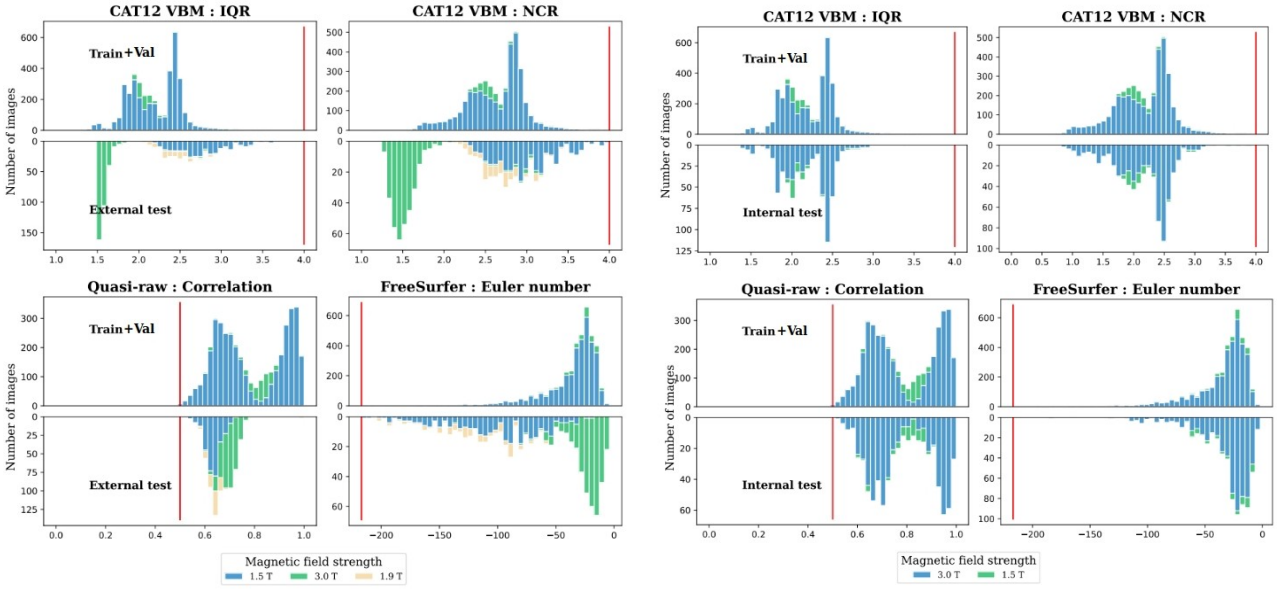
FreeSurfer

Steps: Cortical analysis was performed with FreeSurfer "recon-all" (https://surfer.nmr.mgh.harvard.edu). The analysis stream includes intensity normalization, skull stripping, segmentation of GM (pial) and WM, hemispheric-based tessellations, topology corrections and inflation, and registration to the "fsaverage" template. From the available morphological measures, the Desikan [75] and Destrieux [98] ROI-based cortical thickness (CT), surface area (SA), and curvature (CR) are shared. Specifically, 7 ROI-based features computed both on Desikan and Destrieux atlases are shared including: the cortical thickness (mean and standard deviation), GM volume, surface area, integrated mean and Gaussian curvatures and intrinsic curvature index. Moreover, vertex-wise cortical thickness, curvature and average convexity features [97] (measuring the depth/height of a vertex above the average surface) are also accessible on the high-resolution seven order icosahedron. To allow inter-hemispheric cortical surface-based analysis, we further transform the right hemisphere features into the left one, using the symmetric "$fsaverage\_sym$" Freesurfer template and the "xhemi" routines [119]. The final vertex-wise cortical features comprise $163,842$ nodes per hemisphere.

Quality control: Similarly with quasi-raw and VBM, we first performed a visual analysis on images ranked by the correlation score. In addition we used the Euler number as a metric of quality and we retained images at a threshold greater than -217, as specified in [237] (see Fig. 4.5).

4.2.3 Train-validation-test splits of OpenBHB with external test for the OpenBHB challenge

For the proposed OpenBHB Challenge (see hereafter section 4.3), we have carefully designed a train-validation-test split (Tab. 4.2) such that the public *training* and *validation* sets and the so-called *internal test* set, that are all issued from OpenBHB, share the essential statistical properties needed for the challenge, i.e., similar age, sex, and site distributions. Additionally, to assess the generalization powers of submitted models in the challenge, we have built an independent *external test* set (issued from other sites) described hereafter. Figure 4.7 gives a



(a) Train∪Val (OpenBHB) + External Test

(b) Train∪Val (OpenBHB) + Internal Test (OpenBHB)

Figure 4.5: OpenBHB quality assessment and image selection. Several metrics of quality have been used to perform the quality check (QC) on OpenBHB (train+internal test) and the external test. IQR: Image Quality Rating; NCR: Noise Contrast Ratio; Correlation: Average correlation between each registered images and all the other ones (Z-transformed). The manually determined cutting threshold used as inclusion criteria in QC is indicated by the red vertical line. A clear domain gap is observed on VBM and SBM (FreeSurfer) data between train and external test. However we have similar image quality between OpenBHB train and internal test for all modalities (VBM, SBM, Quasi-raw).

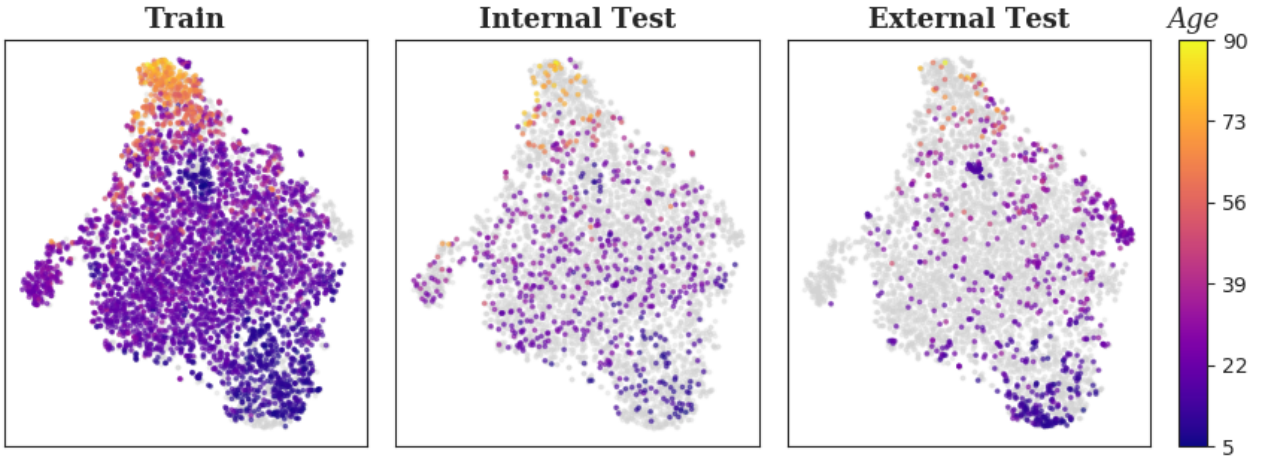


Figure 4.6: t-SNE representation of VBM ROI normalized by TIV for each sample in the challenge splits. Internal test is representative of the training+validation sets in terms of covered regions while external test has regions not represented in train (especially for younger participants, bottom and right regions).

quick visual overview of the splitting strategy. Notably, both internal and external tests are currently kept private to the participants in the context of the OpenBHB Challenge.

Training, Validation and *Internal* test splits of OpenBHB. We used a stratified sampling of OpenBHB similar distributions of essential variables, i.e., age, sex, and site. Consider-

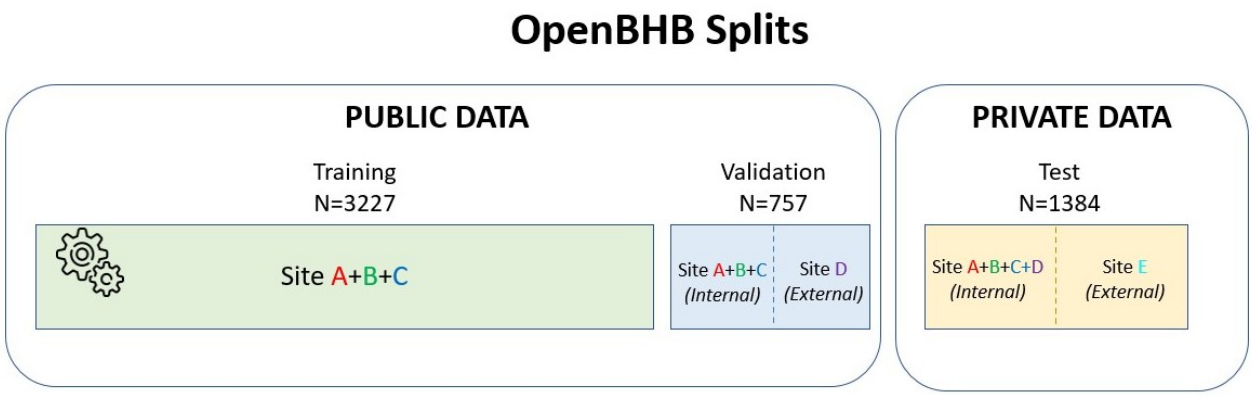


Figure 4.7: Splitting strategy used for OpenBHB. The public data available are split into a training and validation set (useful for cross-validation and to derive comparable public results). Private data (used to score the models submitted to the OpenBHB challenge) are composed of 2 subsets: an internal test (stratified on age, sex and site from the OpenBHB dataset) and an external test (independent from OpenBHB and with acquisition centers distinct from the public data). Importantly, the validation set is built in a similar fashion from public OpenBHB data to allow participants to derive all the challenge's metrics.

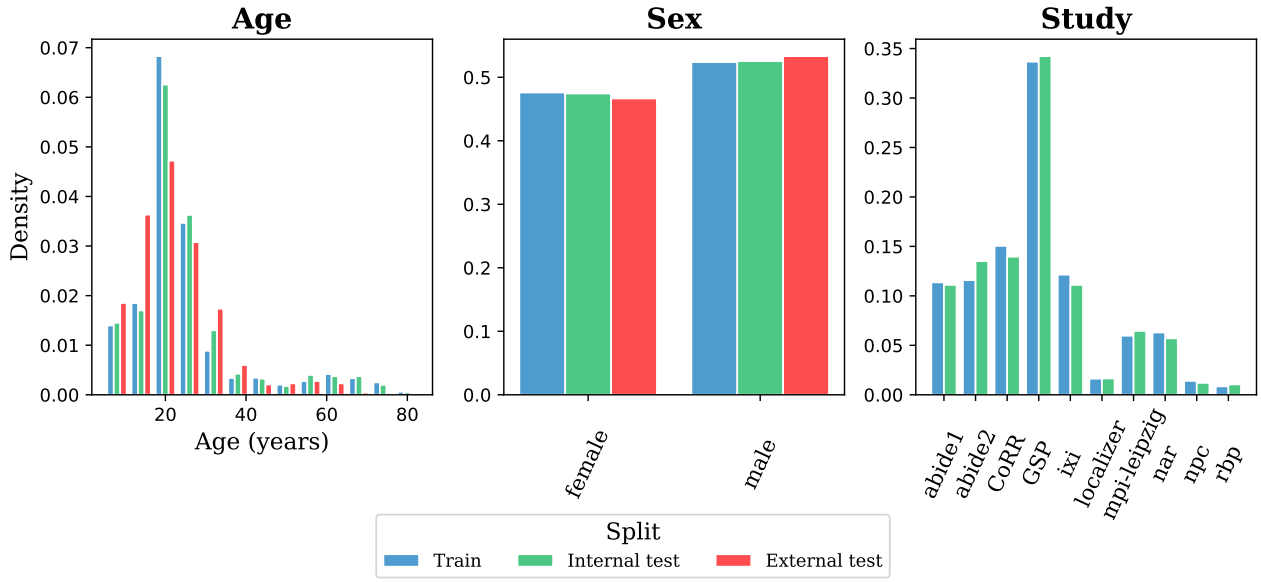


Figure 4.8: Age, sex and study distribution between the training+validation and internal testing splits defined for the OpenBHB Challenge. All statistics are well preserved between both splits, thus avoiding any obvious bias in ML predictions.

ing the large number of sites (> 60) and age bins in OpenBHB - for a reasonable binarization scheme (e.g., 5-years bins) - it is prohibitive to use the naive stratification approach based on the cartesian product between sites, binarized age, and sex (there would not be enough samples per bin). To properly define such a split, we used the iterative stratification algorithm [252] that tries to optimally preserve the different age, sex, and site histograms between train and test. Using this method, we obtained a training/internal test split that holds well age, sex, and site statistics (see Fig. 4.8 and Table 4.2).

Split	Characteristic	# Subjects	Age	Sex (%M)	# Sites×Acq
Train	Public, from OpenBHB	3227	25.2 ± 14.6	52	58
Validation	Public, from OpenBHB	757	23.8 ± 12.8	55	64
Internal Test	Private, from OpenBHB	664	25.3 ± 14.2	52	64
External Test	Private, new sites	720	22.3 ± 11.1	53.3	8

Table 4.2: Training, Validation, Internal (stratified) and External test splits for the OpenBHB Challenge. 679 images from OpenBHB have missing label information (e.g acquisition setting) and are not considered in this challenge. External test is **fully independent** from OpenBHB and it is only used to assess the generalization capacity of submitted algorithms in the OpenBHB Challenge.

External test. In the context of the OpenBHB Challenge, we also built an external test, fully independent from OpenBHB that also preserves age and sex statistics (see Fig. 4.8 and Table 4.2) but that contains images acquired on independent sites (*i.e.*, with no site overlap with OpenBHB). These sites are kept private to avoid any bias or data leakage during the challenge. The external test set is used to evaluate the generalization capacity of ML models on never-seen sites for brain age prediction (see section 4.3 for more details). As for the OpenBHB challenge, MRI scans from this external test have been acquired on both 1.5T and 3T in geographically distinct locations, across 8 acquisition centers.

Region coverage analysis between training, validation and tests Previously we showed that training, validation and the 2 test sets share the same age and sex distributions (along with the same site distribution between training and internal test). As we saw in Fig. 4.2, we can use t-SNE visualization tool on VBM data to analyze the appearance similarity between training and internal or external test sets (similarly with [45] in the context of semi-supervised learning for labelled/unlabelled splits). Specifically, we performed dimensionnality reduction with t-SNE to project ROI VBM data in a 2D space. Then we analyzed the regions covered by training, internal test and external test in this space. In Fig. 4.6, we observe that most regions are covered both by the training and internal test in the t-SNE projected space while some regions are only covered by the external test (e.g bottom and right regions, usually for younger participants). This suggests that the training set covers well the internal test but there is a domain gap between training and external test. To quantitatively assess the coverage of both internal and external tests over training, we took inspiration from [45] by defining the Intersection Over Union (IoU) metric between train and test. We have defined the regions covered by each split by performing a one-class SVM algorithm on data reduced by t-SNE and we have computed IoU between the regions covered by 2 splits (see [45] for more details). In our case, we have IoU=0.94 between {training,validation} and internal test and IoU=0.64 between {training,validation} and external test. This further supports our claim that our training and validation sets represent well our internal test set while the external test covers new regions, distinct from training and validation. This could be directly related to site effect that is present in our dataset (see section 4.3.4) and, thus, we favor algorithms insensitive to domain gap in

the OpenBHB challenge.

4.2.4 Data organization and accessibility

Data sharing. A data sharing platform distributes the prepared OpenBHB dataset. All up-to-date information are centralized at this location https://baobablab.github.io/bhb.

Data organization. Currently, we are only sharing the training and validation splits for the OpenBHB challenge, in order to avoid data leakage. All modalities (quasi-raw, VBM, SBM) are stored in NumPy (.npy) format to allow easy cross-platform implementation. A resource folder contains information about the actual geometry of all manipulated data (i.e VBM and quasi-raw MNI templates as well as ROI labels for the Desikan [75] and Destrieux [98] atlases used by FreeSurfer and for the Neuromorphometric atlas used by CAT12). The data are stored in N-dimensional arrays, where the first two dimensions are sample size and number of modalities (or channels, equal to 1 here), followed by the data dimension. All metadata are stored in the *participants* TSV file and quality control metrics in the *qc* TSV file. Finally, the 64 pair (site, acquisition setting) labels used in the OpenBHB challenge are available in the TSV file *official_site_class_labels*. In more details, the directory contains:

- **participants.tsv**: metadata table with columns participant identifier, study, sex, age, site, acquisition settings, TIV, CSFV, GMV, and WMV (cf. Section 4.2.1).
- **qc.tsv**: quality control table with columns participant identifier, recon-all Euler, CAT12 VBM NCR, CAT12 VBM IQR, and quasi-raw correlation (cf. Section 4.2.2).
- **official_site_class_labels.tsv**: MRI images are biased according to the specific scanner used *and* the acquisition protocol set (e.g., repetition time, echo time, etc.). Consequently, the correct confounding variable to remove is the *pair* (site, acquisition setting). As a result, we discretized all these pairs to create the confounding variable "**siteXacq**" that is defined in this file. All participants in the OpenBHB Challenge should use this variable to remove "site effect". Thus, when referring to "site removal", we implicitly englobe both the scanner and acquisition protocol used to generate the final MRI image.
- **sub-*_desc-gm_T1w.npy**: the GM VBM image with shape $[1 \times 1 \times 121 \times 145 \times 121]$ (see Section 4.2.2). The first two dimensions represent the number of sample and channel (only one sample and GM tissue here) while the last three dimensions are the spatial image dimension ($1.5mm^3$ spatial resolution registered on MNI template). The corresponding MNI template is stored as NIfTI in *resource/cat12vbm_space-MNI152_desc-gm_TPM.nii.gz*.
- **sub-*_preproc-cat12vbm_desc-gm_ROI.npy**: the GM volumes averaged on the anatomical Neuromorphometrics template with shape $[1 \times 1 \times 284]$ (see Section 4.2.2). This template includes 142 cortical and sub-cortical regions for both GM and CSF volumes, thus

totalizing 284 GM volumes that are stored in the last dimension. The template can be found as NIfTI file in *resource/neuromorphometrics.nii*.

- **sub-*_preproc-quasiraw_T1w.npy**: the T1w quasi-raw (minimally preprocessed) image with shape $[1 \times 1 \times 182 \times 218 \times 182]$ (see Section 4.2.2). The resulting image contains all brain tissues and it is registered to the MNI template with $1mm^3$ spatial resolution. The last three dimensions are the spatial image dimension. The template can be found as NIfTI in *resource/quasiraw_space-MNI152_desc-brain_T1w.nii.gz*.

- **sub-*_preproc-freesurfer_desc-desikan_ROI.npy**: cortical thickness (with standard deviation), GM volume, surface area, integrated mean (and Gaussian) curvature and intrinsic curvature index averaged on the Desikan cortical template [75] with shape $[1 \times 7 \times 68]$ (see Section 4.2.2). These 7 features are stored in the second (channel) dimension for all 68 brain regions defined by Desikan template (34 by hemisphere), stored in last dimension. Brain region labels and channels order can be found in *resource/freesurfer_atlas-desikan_labels.txt* and *freesurfer_channels.txt* respectively.

- **sub-*_preproc-freesurfer_desc-destrieux_ROI.npy**: same cortical thickness, GM volume, curvature and surface area measures as previously, averaged on the Destrieux cortical template [98] with shape $[1 \times 7 \times 148]$ (see Section 4.2.2). This template includes 148 brain regions (74 by hemisphere), on which the same 7 features are computed. Brain region labels and channels order can be found in *resource/freesurfer_atlas-destrieux_labels.txt* and *resource/freesurfer_channels.txt* respectively.

- **sub-*_preproc-freesurfer_desc-xhemi_T1w.npy**: cortical thickness, curvature, average convexity features [97] and Desikan cortical parcellation measures computed on the high-quality "fsaverage" mesh with shape $[1 \times 8 \times 163842]$ (see Section 4.2.2). Both right and left hemisphere measures are provided on the "$fsaverage\_sym$" template (163842 vertices), providing $4 \times 2 = 8$ features on each vertex that are stored in second (channel) dimension, the last dimension being the mesh. Channels order can be found in *resource/freesurfer_xhemi_channels.txt*

4.3 OpenBHB challenge: representation learning for age prediction with site effect removal

We propose a challenge to compare the models capacity to encode a relevant representation of the data that preserves the biological variability associated with age while removing the site-specific information. All required information, data loading, models submission, etc. is described on the web page: https://baobablab.github.io/bhb/challenges/age_prediction_with_site_removal.

4.3.1 Background

Predicting phenotype (such as age and sex) from brain imaging data is a key challenge to answer several exciting questions (e.g., biomarker discoveries for psychiatric disorder or neurocognitive impairment with brain age, personalized medicine with normative modeling, etc.). Recent efforts in ML/DL for neuroimaging along with the availability of large-scale datasets have led to the development of models increasingly accurate, capable of predicting chronological age within 2-3 years, and sex with 99% accuracy from brain MRI. However, these models are severely biased by non-biological variability, particularly associated with acquisition sites (different acquisition protocols, manufacturers, magnetic fields, etc.). These sources of variability inevitably limit the performance of such models and can even bias the neuroimaging community towards over-optimistic results.

4.3.2 Challenge description

Consequently, we propose a representation learning challenge using the OpenBHB dataset. The aim is to learn a representation of the 3D T1 anatomical MRI pre-processed with the three pipelines described previously that full-fills 2 key properties: 1) the representation should be predictive of biological age; 2) site information should be removed from the representation. Thus, we aim to compare the capacity of the proposed models to encode a relevant representation of the data (feature extraction and dimensionality reduction) that preserves the biological variability associated with age while removing the site-specific information. The algorithms submitted must output a vector of features with $p < 10^4$ dimensions for each input data. Both quasi-raw, VBM, and SBM features (including ROI-based and voxel-wise features for VBM and SBM data, see Section 4.2.4) will be given as input data and the participants are free to use only a subset of these modalities (for instance only voxel-wise VBM and ROI-based SBM).

Models submitted will be evaluated with the standard linear evaluation protocol (detailed hereafter) to predict age and site from the embedded features. A general overview of the model evaluation workflow for this challenge is depicted in Fig. 4.9.

Linear evaluation details. A logistic regression is used to evaluate the representation quality on site prediction (as an innovative way to measure if site information has been removed). It is trained on the public data encoded by the model submitted and tested on the private data (only internal test) also encoded. As for age prediction, a Ridge regression is trained on the public data and tested on both internal and external test sets (see section 4.2.3). To generate more robust metrics for this challenge, a 3-fold CV scheme has been implemented for the training phase of the linear probe (see Fig. 4.9). This 3-CV scheme is implemented on the whole public data-set (training+validation). As a result, for each metric, a mean and standard deviation are computed and reported in the final official leaderboard.

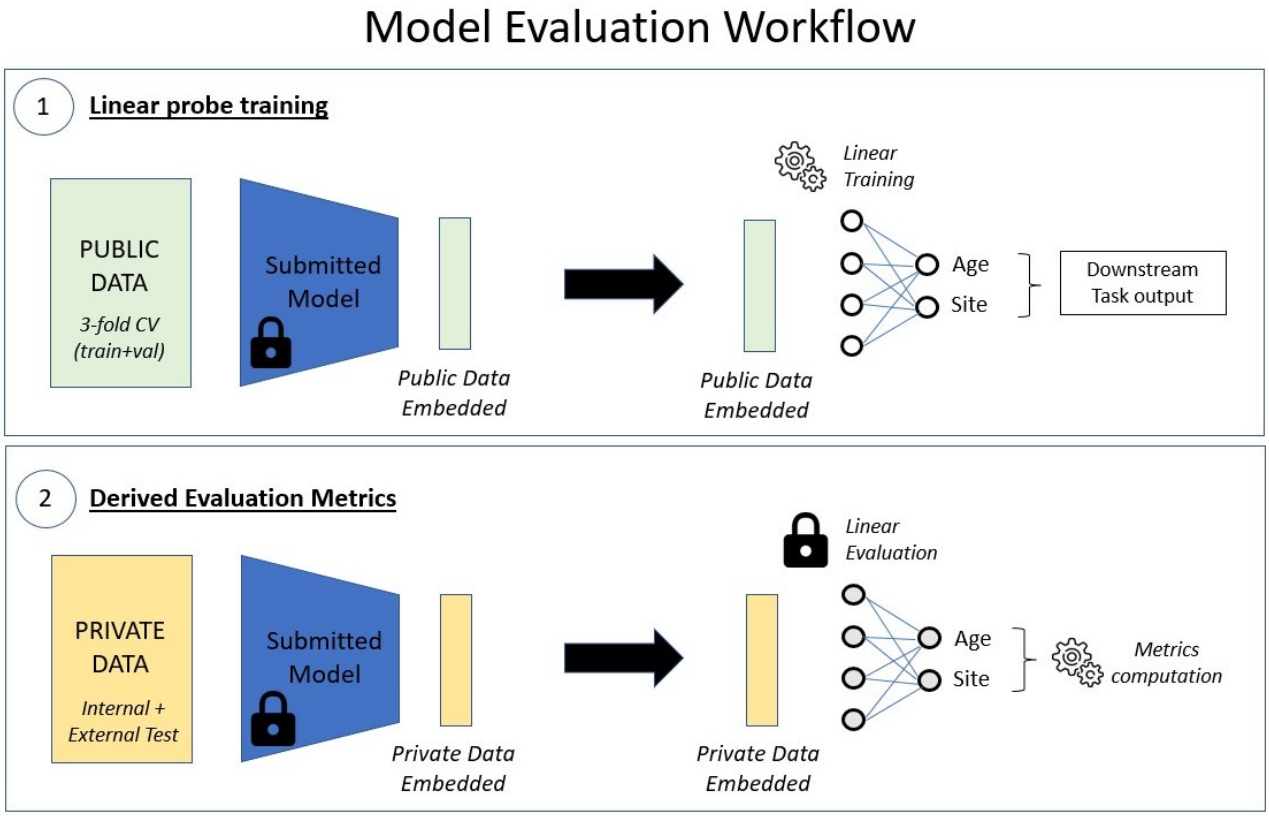


Figure 4.9: Model evaluation workflow of a new submission. When a new trained model is submitted to our platform, a linear probe (regressor for age prediction and classifier for site classification) is trained on top of the public embedded data (i.e. public data encoded by the submitted model). Once trained, this linear probe predicts the downstream targets (age and site) on the private embedded data (age is predicted from both private internal and external tests while site is predicted from private internal test only). 3 metrics are then derived: Mean Absolute Error (MAE) for age prediction on internal and external test; Balanced Accuracy (BAcc) for site prediction on internal test. These 3 metrics are combined to derive the final challenge metric $\mathcal{L}_c$ (see eq. 4.1).

Metric. We have developed a novel metric that jointly evaluates two critical properties of the learned representation: its robustness w.r.t sites and the quantity of information preserved w.r.t chronological age. This metric combines two reference metrics: Mean Absolute Error for age prediction (MAE, to be **minimized**), and Balanced Accuracy for site prediction (BAcc, it should be equal to random chance). BAcc is the preferred metric for classification since sites distribution is heavily imbalanced in OpenBHB (see Fig. 4.8).

To compute the challenge's metric described above, 2 distinct test sets has been derived (see Section 4.2.3): (i) an *Internal test* set containing images from the same sites as the training set (both in OpenBHB); (ii) an *External test* completely independent of OpenBHB, with no site overlap. Both sets will be used to compute MAE for age prediction. Only *Internal test* is used to derive BAcc for site prediction. All metrics will be displayed during the challenge, and the overall ranking will be based on the following metric (the lower, the better):

$$\mathcal{L}_c = \text{BAcc}(\text{sites})^{0.3} \cdot \text{MAE}_{\text{external test}}(\text{age}) \tag{4.1}$$

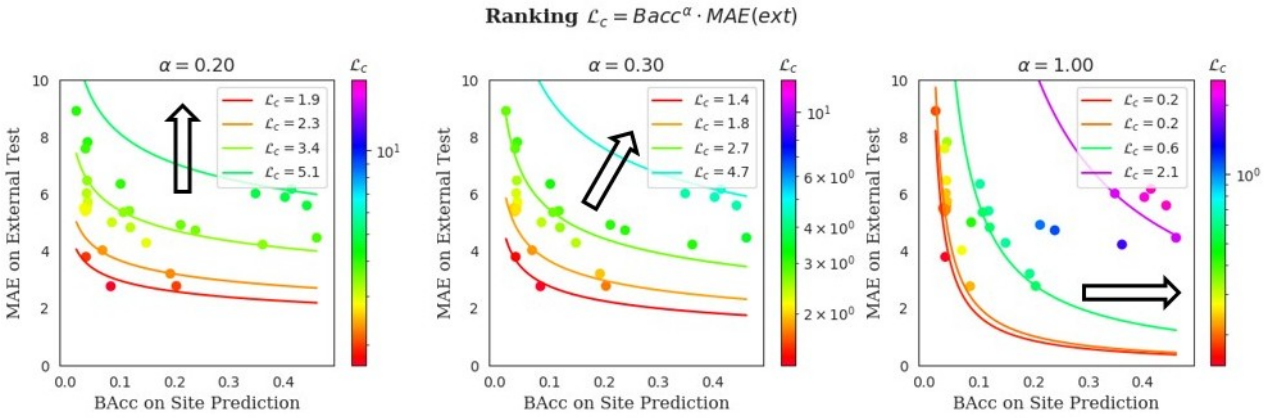


Figure 4.10: Optimal challenge metric for ranking submitted algorithms during the OpenBHB challenge. Each point represents an algorithm performance (CNN or MLP) ran on either VBM, Quasi-Raw or SBM data with a specific architecture in the (MAE(age), Bacc(sites)) plane. Color represents the ranking. Perfect algorithms should be in the bottom left corner. Isoline with constant $\mathcal{L}_c$ (i.e same ranking) is represented with solid colored lines. The following metric is tested: $\mathcal{L}_c = Bacc^{\alpha} \cdot MAE(ext)$ with α a hyper-parameter. Black arrows represent the decreasing ranking trend of the submitted algorithms (e.g. for $\alpha = 1$, algorithms with low Bacc(sites) have good ranking, no matter their MAE(age) while for $\alpha = 0.2$, algorithms with low MAE(age) have good ranking, no matter their Bacc(sites)). $\alpha = 0.3$ is a good empirical trade-off to have good ranking of algorithms that both i) preserve age variability (low MAE) and ii) remove site information (low Bacc). The ranking metric $\mathcal{L}_c = BAcc^{0.3} \cdot MAE(ext)$ is the final choice retained for the OpenBHB challenge.

If the representation learnt retains all site information, then the BAcc is equal to 1. The other way around, if the representation learnt is independent of the site, then the BAcc is no more than random chance which is $1/N_{sites}$ where N_{sites} is the number of sites ($N_{sites} = 64$ for this challenge). As a result, BAcc and MAE give complementary information to quantitatively assess whether site information is completely removed from the representation (low BAcc) and biological variability is preserved (low MAE). One remaining question is: what weight α do we chose as a trade-off for ranking algorithms ? In Fig 4.10, we have performed an analysis of several baseline algorithms (CNN and MLP) trained with several modalities (VBM, SBM, Quasi-Raw) and we have represented their performance in the 2D plane (MAE(ext), Bacc(sites)). Based on this analysis, we have selected the optimal weight $\alpha = 0.3$ such that i) algorithms are ranked according to MAE in priority ($\alpha \ll 1$) but ii) models with similar MAE are ranked through BAcc ($\alpha > 0$). In particular, it ensures that perfectly debiased algorithms that poorly predict age have poor ranking, as well as good age predictors with very strong biased representations towards sites.

The two test sets are hidden to the participants, and they only have access to the training and validation sets with all meta-data information described in section 4.2.1 (in particular, participant's age and sex and acquisition site).

4.3.3 Leaderboard and submission

The leaderboard (available here) contains the 3 metrics defined in the current challenge: MAE for age prediction (computed on internal and external test), BAcc for site prediction (computed only on internal test) and $\mathcal{L}_c$ (as defined in Eq. 4.1). The algorithms are ranked according to this last metric. The full submission process (including the expected code) is described in the challenge web page.

Impact in neuroimaging. This challenge tackles several key problems encountered by the neuroimaging community that require the development of new innovative algorithms that might be borrowed from the computer vision field. Our three preprocessing pipelines allow to use both DL models and standard linear or kernel methods (the latter working on VBM and SBM data while the former works on all modalities). Furthermore, site can be viewed as a confounding variable to remove so this challenge encourages the use of debiasing algorithms[17, 21, 41, 62, 273], a hot topic in computer vision often related to Trustworthy AI and Fairness. We envisage an increasing interest by the vision community in this field (e.g., aiming at building racially or gender debiased models). This challenge also tries to build upon new evaluation strategies that have emerged in representation learning with DL, in particular the linear evaluation setting[5, 43, 52, 136]. We hope it urges researchers to develop new methods that can be translated into an unsupervised setting (such as new self-supervised regularization terms), and thus enhance the DL capacity to generalize well to different tasks (such as brain tumor segmentation or computer-aided diagnosis). Finally, generalizing well on data from never-seen sites implies a good robustness of the algorithm developed to out-of-domain images (related to Domain Adaptation, also a hot topic in the computer vision field-see for instance [296] for a comprenhensive survey). Consequently, this challenge intents to bring together both neuroimaging and computer vision communities on a new large-scale 3D biomedical dataset.

4.3.4 Name-that-site performance

In order to assess the current bias in OpenBHB, we played at the game "Name-That-Site", inspired by [279] on natural images and [293] on brain MRI (originally created as "Name-That-Dataset"). We train a classifier on different input data x to classify between the 64 pairs (site, acquisition setting) in the OpenBHB Challenge and we test it by following the challenge training+validation/test splits defined section 4.2.3. We used a CNN for VBM and Quasi-Raw inputs and linear logistic regressions for ROI surface-based measures and age.

Training details

We considered 3D adaptation of three representative CNN architectures: AlexNet inspired from [2] (5 convolutional layers and 2.5M parameters), ResNet18 [135] (18 convolutional layers and 33.2M parameters) and DenseNet121 [149] (121 convolutional layers and 11.3M parame-

ters) for whole-brain imaging data (VBM and quasi-raw). Their implementation is available on challenge web page : https://baobablab.github.io/bhb/challenges/age_prediction_with_site_removal. We trained them for 300 epochs with an initial learning rate $\alpha = 10^{-4}$ reduced by factor 0.9 every 10 epochs. We used the standard cross-entropy loss optimized with Adam [172] optimizer $(\beta_1, \beta_2) = (0.9, 0.999)$. As for SBM data (both ROI-based and mesh-based features), we simply optimized a logistic regression using scikit-learn [217] and we cross-validated the regularization term $C \in \{0.01, 0.1, 1, 10, 100\}$ on the validation set. All SBM features are flattened (i.e we merge channels and spatial dimensions) to perform logistic regression.

Results

Results in Table 4.3 first indicate that all modalities preserve site information, especially input data in their rawest form (Quasi-Raw) with 83% BAcc. Second, age give small hints about site (2.86% BAcc) indicating that participants inclusion is biased for some datasets (*e.g.,* ABIDE) but it remains negligible compared to other modalities (always $>$ 38% BAcc). These results are in line with recent literature [293] and it highlights the high non-biological heterogeneity that remains in MRI images even after non-linear registration and normalization (e.g VBM). Finally, these results suggest that Bacc is an appropriate metric to quantify the bias in models representation.

Input	**Model**	**Balanced Accuracy**(%)	**Accuracy**(%)
Quasi-Raw	AlexNet [181]	83.1	86.8
	DenseNet [149]	87.8	91.4
	ResNet [135]	79.1	88.7
VBM	AlexNet [181]	42.3	47.8
	DenseNet [149]	62.2	72.4
	ResNet [135]	60.5	71.4
SBM(Xhemi)	Linear	38.0	57.5
SBM(ROI-Destrieux)	Linear	70.5	74.3
SBM(ROI-Desikan)	Linear	67.6	73.9
Age	Linear	2.86	26.7
Random Level	-	1.56	7.35

Table 4.3: Performance for predicting one of the 64 pairs (site, acquisition setting) included in the OpenBHB Challenge. All CNN are 3D adaptation of the original architectures and AlexNet corresponds to the 3D version proposed in [2].

4.3.5 Baselines for the OpenBHB challenge

Next, we performed baseline experiments on different data modalities (Quasi-Raw, VBM, SBM) for the OpenBHB Challenge. Specifically, we trained CNNs with various architectures on 3D whole-brain imaging data (voxel-wise VBM and Quasi-Raw). The objective function is a simple ℓ_1 loss on age prediction for all models.

We tested data-based debiasing model with the popular ComBat [101] residualization method, applied only on training set (both internal and external test sets are left non-harmonized since age and site labels are not available). This approach is compared to the standard brain age prediction DL-based model which does not take site information into account.

Training and evaluation details

We used the same CNN architectures as described Section 4.3.4 for VBM and quasi-raw data. We optimized the ℓ_1 loss between true and predicted age with Adam optimizer and initial learning rate $\alpha = 10^{-4}$ decreased by a factor 0.9 every 10 epochs. Then, the last fully-connected (FC) layer is removed and the representation is evaluated using 1) Logistic Regression for site prediction with cross-validation of ℓ_2-regularization parameter $C \in \{0.01, 0.1, 1, 10, 100\}$ and 2) Ridge Regression for age prediction on internal and external test with cross-validation of regularization in $\{0.01, 0.1, 1, 10, 100\}$. Evaluation procedure is detailed in Section 4.3.2 and it is executed in the same manner for all submitted models on server side. These models have been submitted to the official challenge and the results can also be found in the official leaderboard.

ComBat residualization is performed only on training data using age, sex (biological variables to keep) and site (confounding variable to remove). We used the official GitHub implementation of ComBat[2].

For SBM ROI-based and mesh-based data, we used vanilla Multi-Layer Perceptrons (MLP) with varying depth but constant latent space dimension (128) and number of neurons per hidden layer (128). FC layer is added on top of MLP encoder to optimize ℓ_1 loss on age prediction with a cross-validation on learning rate $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. The evaluation procedure is the same as previously. Importantly, for MLP training on SBM Xhemi (mesh-based measures) we did not include the channels associated to Desikan cortical parcellation (since it does not bring more anatomical information).

Results

We reported the 3 metrics used in the OpenBHB Challenge: MAE on internal and external test and Bacc on site prediction (see section 4.3). The latent space dimension varying across CNN architectures is systematically reported.

First, we notice that all models retain site information without any debiasing strategy. Overall, Quasi-Raw data are more biased than VBM, as we saw previously (see Table 4.3), and CNN preserve this bias to some extent. This is especially true for DenseNet, the best performing network on both VBM and Quasi-Raw on the internal test, one of the models that retains the most site information (8.0% Bacc and 15.2% Bacc on VBM and Quasi-Raw respectively). This would explain the drop in performance on the external test (+4.58 MAE on VBM). In this regard, ResNet is the best trade-off (with also the best ranking for the challenge) as it is robust to site and it generalizes well on the external test. These results also suggest

[2] https://github.com/Jfortin1/neurocombat_sklearn

Method	Model (Latent Dim. # params)	VBM				Quasi-Raw			
		Int. Test		Ext. Test	$\mathcal{L}_c \downarrow$	Int. Test		Ext. Test	$\mathcal{L}_c \downarrow$
		MAE↓	BAcc↓	MAE↓		MAE↓	BAcc↓	MAE↓	
Baseline	DenseNet (1024, 11.2M)	$2.55_{\pm 0.009}$	$8.0_{\pm 0.9}$	$7.13_{\pm 0.05}$	3.34	$2.48_{\pm 0.03}$	$15.2_{\pm 0.6}$	$2.92_{\pm 0.07}$	1.66
	ResNet (512, 33.2M)	$2.67_{\pm 0.05}$	$6.7_{\pm 0.1}$	$4.18_{\pm 0.01}$	1.86	$2.60_{\pm 0.003}$	$7.6_{\pm 0.1}$	$2.85_{\pm 0.004}$	1.31
	AlexNet (128, 2.5M)	$2.72_{\pm 0.01}$	$8.3_{\pm 0.2}$	$4.66_{\pm 0.05}$	2.21	$2.96_{\pm 0.005}$	$16.2_{\pm 0.5}$	$3.65_{\pm 0.009}$	2.11
ComBat [101]	DenseNet (1024, 11.2M)	$5.92_{\pm 0.01}$	$2.23_{\pm 0.06}$	$10.48_{\pm 0.17}$	3.38	N.A.	N.A.	N.A.	N.A
	ResNet (512, 33.2M)	$4.15_{\pm 0.009}$	$4.5_{\pm 0.0}$	$4.76_{\pm 0.03}$	1.88	N.A.	N.A.	N.A.	N.A
	AlexNet (128, 2.5M)	$3.37_{\pm 0.01}$	$6.8_{\pm 0.3}$	$5.23_{\pm 0.12}$	2.33	N.A.	N.A.	N.A.	N.A

Table 4.4: Baselines obtained with 1) no de-biasing strategy (first 3 rows) and 2) ComBat residualization on training data (last 3 rows) with VBM and Quasi-Raw data for 3 representative CNN families. MAE=Mean Absolute Error and BAcc=Balanced Accuracy (in %). Balanced Accuracy on internal test should be compared to baseline $BAcc = 2.86\%$ which is the prediction power of true age for predicting site in the internal test set.

that having a deeper network (e.g., DenseNet and 121 layers) does not translate necessarily in better generalization performance (e.g., compared to ResNet with 18 layers), in line with [86].

Interestingly, ComBat harmonization does remove most of site bias in CNN representation space (with site prediction Bacc decreased by 2% but not still matching the one obtained using only age as input, see Table 4.3). However, it also heavily degrades CNN performance on age prediction for all testing sets (in particular for DenseNet). ComBat is not fitted for Quasi-Raw data as it mainly relies on voxel-wise statistics, and raw data are not properly registrated voxel-wise across images. Consequently, we did not evaluate this approach on Quasi-Raw images. For completeness, we finally evaluated several vanilla Multi-Layer Perceptrons (MLP) with

Modality	Model (Hidden Layers)	Baseline				ComBat [101]			
		Int. Test		Ext. Test	$\mathcal{L}_c \downarrow$	Int. Test		Ext. Test	$\mathcal{L}_c \downarrow$
		MAE↓	BAcc↓	MAE↓		MAE↓	BAcc↓	MAE↓	
FSL-Xhemi	*Linear*(Original)	4.69	38.0	5.96	4.46	5.08	30.8	6.19	4.35
	MLP(128)	$3.72_{\pm 0.04}$	$\mathbf{8.6_{\pm 1.0}}$	$5.31_{\pm 0.07}$	**2.54**	$4.06_{\pm 0.05}$	$6.00_{\pm 0.05}$	$5.0_{\pm 0.2}$	2.15
	MLP(128, 128)	$3.90_{\pm 0.03}$	$16.4_{\pm 1.14}$	$5.06_{\pm 0.06}$	2.94	$4.49_{\pm 0.04}$	$6.6_{\pm 0.4}$	$5.34_{\pm 0.03}$	2.36
	MLP(128, 128, 128)	$3.75_{\pm 0.008}$	$16.7_{\pm 1.28}$	$5.07_{\pm 0.07}$	2.96	$\mathbf{4.27_{\pm 0.02}}$	$4.2_{\pm 0.1}$	$5.43_{\pm 0.04}$	2.1
	MLP(128, 128, 128, 128)	$\mathbf{3.49_{\pm 0.01}}$	$14.4_{\pm 1.05}$	$\mathbf{4.85_{\pm 0.02}}$	2.71	$4.28_{\pm 0.01}$	$\mathbf{4.0_{\pm 0.04}}$	$\mathbf{5.25_{\pm 0.01}}$	**2.0**
FSL-ROI Destrieux	*Linear*(Original)	4.96	70.5	7.21	6.49	6.44	36.7	9.30	6.88
	MLP(128)	$3.96_{\pm 0.01}$	$24.9_{\pm 0.9}$	$6.72_{\pm 0.12}$	4.43	$5.65_{\pm 0.08}$	$15.0_{\pm 1.19}$	$11.79_{\pm 0.98}$	6.67
	MLP(128, 128)	$3.36_{\pm 0.01}$	$32.4_{\pm 0.94}$	$4.89_{\pm 0.10}$	3.49	$4.52_{\pm 0.04}$	$19.8_{\pm 0.9}$	$6.27_{\pm 0.19}$	3.86
	MLP(128, 128, 128)	$\mathbf{3.07_{\pm 0.02}}$	$21.1_{\pm 1.13}$	$4.69_{\pm 0.05}$	2.94	$4.39_{\pm 0.02}$	$7.1_{\pm 0.2}$	$5.81_{\pm 0.09}$	2.63
	MLP(128, 128, 128, 128)	$3.12_{\pm 0.005}$	$\mathbf{12.8_{\pm 0.5}}$	$\mathbf{4.48_{\pm 0.02}}$	**2.42**	$\mathbf{4.37_{\pm 0.03}}$	$\mathbf{6.5_{\pm 0.06}}$	$\mathbf{5.39_{\pm 0.15}}$	**2.37**
FSL-ROI Desikan	*Linear*(Original)	5.27	67.6	6.58	5.85	7.97	7.06	13.5	6.09
	MLP(128)	$4.80_{\pm 0.06}$	$19.8_{\pm 2.3}$	$14.7_{\pm 1.16}$	9.04	$6.01_{\pm 0.02}$	$17.3_{\pm 1.91}$	$25.0_{\pm 3.0}$	14.77
	MLP(128, 128)	$3.46_{\pm 0.02}$	$32.6_{\pm 1.43}$	$5.96_{\pm 0.15}$	4.26	$5.21_{\pm 0.11}$	$21.7_{\pm 2.44}$	$7.00_{\pm 0.44}$	4.43
	MLP(128, 128, 128)	$3.51_{\pm 0.01}$	$22.7_{\pm 1.31}$	$\mathbf{4.95_{\pm 0.01}}$	3.17	$\mathbf{4.58_{\pm 0.02}}$	$9.2_{\pm 0.7}$	$\mathbf{5.19_{\pm 0.03}}$	**2.54**
	MLP(128, 128, 128, 128)	$\mathbf{3.37_{\pm 0.003}}$	$\mathbf{11.2_{\pm 0.5}}$	$5.27_{\pm 0.006}$	**2.73**	$5.23_{\pm 0.006}$	$\mathbf{5.9_{\pm 0.16}}$	$6.02_{\pm 0.13}$	2.57

Table 4.5: Baseline results on SBM data using simple Multi-Layers Perceptron (MLP) with increasing depth (from 1 to 4) trained to predict age. The latent space dimension is fixed to 128 for all models. ComBat residualization is used only during training of the MLP models as a debiasing method. Linear models are also evaluated for comparison purposes (even if they are not accepted as a valid model in this challenge). Overall, deeper models lead to better data representation but they are still very biased by site without any debiasing strategy. ComBat residualization degrades performance for age prediction but it removes efficiently site bias (especially for the deepest models).

varying depth on brain age prediction using surface-based mesh (Xhemi) and ROI features on

2 atlases, namely Destrieux and Desikan (see section 4.2.2 for more details). As opposed to VBM, SBM also includes geometrical properties of brain sulci and gyri (e.g local curvature). Thus, it conveys complementary information that may have been lost in volume-based VBM data.

For comparison purposes, we also added the performance of a *linear* model trained directly on input data. We emphasize this model is not authorized in the official challenge (since we expect a model that outputs a low-dimensional representation of the input data). Nevertheless, it gives a general baseline for all submitted models. This linear model is trained on the whole training+validation set with a cross-validation procedure details in Section 4.3.5. We did not perform a 3-fold CV for training the linear probe (corresponding here directly to the linear model) so we do not report a standard deviation.

In Table 4.5, we observe that a finer atlas with 148 regions (Destrieux) leads to better brain age estimation as opposed to coarse atlas (Desikan with only 68 regions), but it also preserves more site information. Furthermore, the deeper the MLP is, the better in terms of MAE for both atlases. Overall, whole-brain approach seems better suited as it enables CNN to extract fine-grained information for brain age, making more accurate predictions while removing more non-biological site information. Nevertheless, merging VBM with SBM data may result in a better representation as both modalities can bring complementary information to model brain development.

These results further justify the necessity of the OpenBHB Challenge since current standard neuroimaging debiasing methods all have their limitations and it can provide an innovative way to develop and benchmark new ML algorithms on brain age prediction, multi-site harmonization and debiasing.

4.4 A first contrastive learning approach for debiasing

In the previous section, we have shown that current deep models representation is biased by acquisition site when they are trained to predict age with ℓ_1 loss and without any debiasing strategy. We propose here to view the problem from a metric learning point-of-view using a contrastive learning approach. Contrary to traditional data harmonization technique (such as ComBat [101]), this method does not modify input data but rather uses a regularization strategy during training. We introduce the new concepts[3] along with the novel loss before showing first results submitted to the OpenBHB challenge as our use case.

4.4.1 Supervised learning from a metric learning perspective

Contrastive learning setup. Let $x \in \mathcal{X}$ be an original sample (i.e., anchor), $(x_i^+)_{i\in[1..P]}$ a set of similar (positive) samples and $(x_j^-)_{j\in[1..N]}$ a set of dissimilar (negative) samples. In

[3]We recommend reading the section 3.1.2 in Chapter 3 for a general introduction of contrastive learning with presentation of the basic notions.

general, positive samples (x_i^+) can be defined in different ways depending on the problem: using transformations of x (unsupervised setting), samples belonging to the same class as x (supervised) or with similar image attributes of x (weakly-supervised). The definition of negative samples (x_j^-) varies accordingly. Here, we focus on the supervised case where we predict the age. Positive (resp. negative) samples are brain images of subjects with similar (resp. dissimilar) age. Contrastive learning methods look for a parametric mapping function $f : \mathcal{X} \to \mathbb{S}^{d-1}$ that maps "semantically" similar samples close together in the representation space (a $(d-1)$-sphere) and dissimilar samples far away from each other. Once pre-trained, f is fixed and its representation is evaluated on a downstream task, such as age prediction here, through linear evaluation on a test set.

We define $s(f(a), f(b))$ as a similarity measure (e.g., cosine similarity) between the representation of two samples a and b. Please note that since $||f(a)||_2 = ||f(b)||_2 = 1$, using a cosine similarity is equivalent to using a L2-distance $d(f(a), f(b)) = ||f(a) - f(b)||_2^2 = 2 - 2s(f(x), f(b))$. Using an ϵ-margin metric learning point of view [60, 127, 248, 261, 295, 298], probably the simplest contrastive learning formulation is looking for a mapping function f such that the following condition is satisfied:

$$\forall j, \underbrace{s(f(x), f(x^+))}_{s^+} \geq \underbrace{s(f(x), f(x_j^-))}_{s_j^-} + \epsilon \tag{4.2}$$

where $\epsilon \geq 0$ is a margin between positive and negative samples and we consider, for now, a single positive sample. This simple constraint can in fact be re-casted as an optimization problem (using *LogSumExp* approximation of max operator), leading to InfoNCE loss [52, 211]:

$$\arg\min_f \max(-\epsilon, \{s_j^- - s^+\}_{j=1,\dots,N}) \approx \arg\min_f -\log \frac{\exp(s^+)}{\frac{1}{N+1}\left(\exp(s^+ - \epsilon) + \sum_j \exp(s_j^-)\right)} \tag{4.3}$$

In the previous equation, $\epsilon = 0$ gives InfoNCE whereas when $\epsilon \to \infty$ we obtain the InfoL1O loss [222]. There are respectively lower and upper bound of the mutual information between X and X^+.

Supervised contrastive loss. The previous InfoNCE loss only contains one positive sample for multiple negatives. In the supervised setting, we may have multiple positive samples for a given anchor x. Interestingly, only imposing condition (4.2) for all positives is actually not enough to retrieve the popular Supervised Contrastive (SupCon) loss [170]. We must add another non-contrastive constraint on the positive samples $s_t^+ - s_i^+ \leq 0 \quad \forall i, t$. This condition forces all positive samples to collapse to a single point in the representation space, however it does not take into account negative samples. That is why we define it as non-contrastive.

Considering both conditions we derive the following optimization problem:

$$\forall i,j \quad s_i^+ \geq s_j^- + \epsilon \quad \text{and} \quad \forall i, t \neq i \quad s_i^+ \geq s_t^+$$

$$\frac{1}{P}\sum_i \max(0, \{s_j^- - s_i^+ + \epsilon\}_j, \{s_t^+ - s_i^+\}_{t\neq i}) \approx \epsilon - \frac{1}{P}\sum_i \log \frac{\exp(s_i^+)}{\frac{1}{N+P}\left(\sum_i \exp(s_i^+ - \epsilon) + \sum_j \exp(s_j^-)\right)} \tag{4.4}$$

when $\epsilon = 0$ we retrieve exactly SupCon.

Regression case. For this challenge, the target is a continuous value (age) so we can re-use the previous y-Aware contrastive formulation introduced in Chapter 3 to define the positive distribution. More precisely, we introduce a similarity function between age y_1 and y_2 as $w_\sigma(y_1, y_2) = K_\sigma(y_1 - y_2)$ with $K_\sigma(u) \propto \exp\left(-\frac{u^2}{2\sigma^2}\right)$ a Gaussian kernel and we write our y-Aware InfoNCE loss as:

$$\mathcal{L}^y_{InfoNCE} = -\sum_{i=1}^{P} \frac{w_\sigma(y, y_i)}{\sum_{k=1}^{P} w_\sigma(y, y_k)} \log \frac{\exp(s_i^+)}{\frac{1}{N+P}\left(\sum_i \exp(s_i^+) + \sum_j \exp(s_j^-)\right)} \tag{4.5}$$

Here y designates anchor's age and $(y_k)_{k=1}^P$ designate positive samples' age.

4.4.2 Proposed regularization

Satisfying condition (4.2) can generally guarantee good downstream performance, however it does not take into account the presence of biases. A model could therefore take its decision based on visual features, the bias, that are correlated with the target downstream task or very easy to learn but that don't actually characterise it. This means that the same bias features would probably have a worst performance if transferred to a different data-set (e.g., different acquisition settings or image quality). Specifically, in contrastive learning, this can lead to settings where we are still able to minimize the SupCon (or y-Aware InfoNCE) loss, but with degraded classification/regression performance. Here, we propose to add *debiasing constraints* that prevent the use of the bias features within the proposed metric learning approach. Similarly to [205], we employ the notion of *bias-aligned* and *bias-conflicting* samples. In our context, a **bias-aligned sample share the same bias attribute of the anchor**, while a **bias-conflicting sample does not**.

Characterization of bias. We denote positive bias-aligned samples with $x^{+,b}$ and positive bias-conflicting samples with $x^{+,b'}$. Given an anchor x, if the bias is "strong" and easy-to-learn, then a positive and bias-aligned sample $x^{+,b}$ will probably be closer in the representation space than a positive bias-conflicting sample. This is why, even in cases in which condition (4.2) is satisfied, we could still be able to distinguish among bias-aligned and bias-conflicting samples. Hence, we say that there is a bias if we can identify an ordering on the learned representations

such as, for example:

$$\forall i,k,t,j \quad \underbrace{d(f(x),f(x_i^{+,b}))}_{d_i^{+,b}} < \underbrace{d(f(x),f(x_k^{+,b'})}_{d_k^{+,b'}} \tag{4.6}$$

This represents the worst-case scenario, where the ordering is total (i.e., $\forall i,k$). Of course, there can also be cases in which the bias is not as strong, and the ordering may be partial.

Regularization for debiasing. Ideally, we would enforce the conditions $d_k^{+,b'} = d_i^{+,b} \quad \forall i,k$, meaning that every positive bias-conflicting sample should have the same distance from the anchor as any other positive (resp. negative) bias-aligned sample. However, in practice, this condition is very strict, as it would enforce uniform distance among all positive (resp. negative) samples. A more relaxed condition would instead force the distributions of distances, $\{d_k^{+,b'}\}$ and $\{d_i^{+,b}\}$, to be similar. Here, we propose new debiasing constraints for positive samples using either the first moment of the distributions or the first two. Using only the average of the distributions, we obtain:

$$\frac{1}{P_a}\sum_i d_i^{+,b} - \frac{1}{P_c}\sum_k d_k^{+,b'} = 0 \iff \frac{1}{P_c}\sum_k s_k^{+,b'} - \frac{1}{P_a}\sum_i s_i^{+,b} = 0 \tag{4.7}$$

where P_a and P_c are the number of positive bias-aligned and bias-conflicting samples respectively.

Calling the first moments $\mu_{+,b} = \frac{1}{P_a}\sum_i d_i^{+,b}$ and $\mu_{+,b'} = \frac{1}{P_c}\sum_k d_k^{+,b'}$, and the second moments of the distance distributions $\sigma^2_{+,b} = \frac{1}{P_a-1}\sum_i (d_i^{+,b} - \mu_{+,b})^2$, $\sigma^2_{+,b'} = \frac{1}{P_c-1}\sum_k (d_k^{+,b'} - \mu_{+,b-})^2$, and making the hypothesis that the distance distributions follow a normal distribution, we can define a new set of debiasing constraints using the Kullback–Leibler divergence:

$$D_{KL}(\{d_i^{+,b}\}||\{d_k^{+,b'}\}) = \frac{1}{2}\left[\frac{\sigma^2_{+,b} + (\mu_{+,b} - \mu_{+,b'})^2}{\sigma^2_{+,b'}} - \log\frac{\sigma^2_{+,b}}{\sigma^2_{+,b'}} - 1\right] = 0 \tag{4.8}$$

In practice, one could also use their symmetric version ($D_{KL}(p||q) + D_{KL}(q||p)$), namely the Jeffreys divergence.

The proposed debiasing constraint can be easily added to any contrastive loss using the method of Lagrange multipliers. They can thus be seen as a regularization term: $\mathcal{R}^{debias} = D_{KL}(\{d_i^{+,b}\}||\{d_k^{+,b'}\})$. Here, we propose to minimize the following objective function:

$$\mathcal{L} = \mathcal{L}^y_{InfoNCE} + \lambda\mathcal{R}^{debias} \tag{4.9}$$

Regression case. As for y-Aware InfoNCE loss, we propose to use the same similarity function w_σ to weight the first and second moments of the distance distribution such that: $\mu_{+,\cdot} = \sum_i \frac{w_\sigma(y,y_i)}{\sum_k w_\sigma(y,y_k)} d_i^{+,\cdot}$ and $\sigma^2_{+,\cdot} = \sum_i \frac{w_\sigma(y,y_i)}{\sum_k w_\sigma(y,y_k)} (d_i^{+,\cdot} - \mu_{+,\cdot})^2$.

4.4.3 Comparison with other debiasing methods

SupCon [170] It is interesting to notice that non-contrastive conditions in Eq.4.4: $s_t^+ - s_i^+ \leq 0 \quad \forall i, t \neq i$ are actually all fulfilled only when $s_i^+ = s_t^+ \quad \forall i, t \neq i$. This means that one tries to align all positive samples, regardless of their bias b. Nonetheless, this condition is enforced uniformly on all positive samples. On the other hand, our formulation distinguishes bias-aligned and bias-conflicting samples in order to put harder constraints between the representation of these 2 populations during optimization.

EnD [273] Constraint in Eq. 4.7 is very similar to what was recently proposed by [273] with EnD. However, EnD lacks of the further constraint on the standard deviation of the distances, which is given by 4.8. An analytical comparison can be found in Appendix C.1.

BiasCon Authors propose a BiasCon loss, which is similar to SupCon but it only aligns positive bias-conflicting samples. It looks for an encoder f that fulfills:

$$s_j^- - s_i^{+,b'} \leq -\epsilon \quad \forall i, j \quad \text{and} \quad s_p^{+,b} - s_i^{+,b'} \leq 0 \quad \forall i, p \text{ and } \quad s_t^{+,b'} - s_i^{+,b'} \leq 0 \quad \forall i, t \neq i \quad (4.10)$$

The problem here is that we try to separate the negative from only the positive bias-conflicting samples, ignoring the positive bias-aligned samples. This is probably why authors proposed to combine this loss with a standard Cross Entropy.

4.4.4 Preliminary results

We report preliminary results of y-Aware InfoNCE on the OpenBHB challenge in Table 4.6, using VBM data. Our model has been trained here only with y-Aware InfoNCE loss. We compare these results to the official baseline results obtained in Section 4.3.5 with no debiasing strategy and simple optimization with ℓ_1 loss. Models are trained for 1000 epochs using Adam optimizer and an initial learning rate $\alpha = 10^{-4}$ decreased by 0.9 every 10 epochs.

Method	Model (features, params)	Int. Test		Ext. Test	$\mathcal{L}_c$
		MAE	BAcc	MAE	
Baseline	DenseNet121 (1024, 11.2M)	$2.55_{\pm 0.009}$	$8.0_{\pm 0.9}$	$7.13_{\pm 0.05}$	3.34
	ResNet18 (512, 33.2M)	$2.67_{\pm 0.05}$	$6.7_{\pm 0.1}$	$4.18_{\pm 0.01}$	1.86
	AlexNet (128, 2.5M)	$2.72_{\pm 0.01}$	$8.3_{\pm 0.2}$	$4.66_{\pm 0.05}$	2.21
ComBat [101]	DenseNet-121 (1024, 11.2M)	$5.92_{\pm 0.01}$	$2.23_{\pm 0.06}$	$10.48_{\pm 0.17}$	3.38
	ResNet18 (512, 33.2M)	$4.15_{\pm 0.009}$	$4.5_{\pm 0.0}$	$4.76_{\pm 0.03}$	1.88
	AlexNet (128, 2.5MM)	$3.37_{\pm 0.01}$	$6.8_{\pm 0.3}$	$5.23_{\pm 0.12}$	2.33
$\mathcal{L}^y_{InfoNCE}$	ResNet18 (512, 33.2M)	$2.66_{\pm 0.00}$	$6.60_{\pm 0.17}$	$4.10_{\pm 0.01}$	**1.82**

Table 4.6: Comparison between baseline experiments with ℓ_1 loss for age regression and Age-Aware InfoNCE (extension of SupCon to regression).

We currently observe that y-Aware InfoNCE performs better than baseline experiments using

ResNet18. We expect that adding our regularization term during optimization will improve the generalization capacity of our model on age prediction while reducing the bias.

4.5 Conclusions and future works with OpenBHB

4.5.1 Towards transfer learning for computer-aided diagnosis

This challenge is a first step towards building new algorithms for phenotype prediction robust across sites. However, as we previously shown in Chap. 3, it is also possible to leverage such multi-center large-scale dataset to significantly improve classification performance on other hard computer-aided diagnosis (CAD) tasks such as Alzheimer's detection or schizophrenia diagnosis, with Transfer Learning (TL). First, pre-training algorithms to remove site effect is critical for computer-aided diagnosis since most clinical datasets with moderate size ($N > 100$) are multi-site (e.g ADNI[156], ABIDE, SCHIZCONNECT, etc.) and it is known [176, 293] that acquisition site can heavily bias ML models. Second, self-supervised algorithms are attracting more and more attention from the medical imaging community[49, 87, 270, 272, 320, 321] since they are able to leverage large un-annotated dataset to improve performance on several downstream tasks with TL. As a result, OpenBHB provides a way to benchmark algorithms on TL, and potentially catalyze research to find new innovative self-supervised algorithms on brain MRI for CAD.

4.5.2 Towards multi-modal integration for new bio-markers discovery

In the context of brain age prediction as a tool for CAD, sMRI data provide meaningful information about subtle anatomical modifications in cortical and sub-cortical structures (e.g atrophies or hypertrophies). On the other hand, resting-state fMRI and Diffusion Weighted Images (DWI) give hints about functional and structural brain connectivity, that can be altered for patients with psychiatric disorders (e.g autism spectrum disorder, schizophrenia or bipolar disorder [18]). However, little is known about the predictive power of a multi-modal approach combining both sMRI, DWI and resting-state fMRI for brain age modelling or CAD. A future release of the OpenBHB dataset might bring some answers. Almost all datasets included in the current OpenBHB release - IXI and NAR excepted - have rfMRI data available for all participants and only a few of them (IXI, NPC, RBP) also have DWI (see Table 4.1). Consequently, we envision to integrate first the rfMRI data as an additional modality in OpenBHB in order to encourage the development of new ML models for multi-modal integration.

Conclusions and Perspectives

Contributions

Single-subject prediction from brain imaging data is crucial for key clinical applications such as personalized medicine and biomarker discovery. It allows answering basic questions about the neuroanatomical signature underlying brain disorders, opening avenues to understand the neurobiological mechanism and, in the end, advancing towards a therapeutic strategy adapted to each patient.

In this thesis, we have studied the representation capacity of deep learning models to solve single-subject prediction tasks in both large-scale and small-scale brain imaging datasets. In a first supervised approach (Chap. 2), we asked how deep models performance scales compared to "standard machine learning" algorithms (i.e. linear and kernel-SVM) for phenotype prediction and mental disorders classification, as we increase dataset size. This first analysis revealed several shortcomings for deep models, in particular its lack of robustness on cross-site cohorts; its dependency to image pre-processing (differing from its well-known advantage on natural images); the limited usefulness of current data augmentation strategies (both geometrical and noise-injected transformations) and data harmonization methods for site-effect removal. Importantly, we showed similar performance between linear models and non-linear deep neural networks for all clinical tasks in the medium-scale data regime ($n \approx 1$k), and a small but significant advantage of the latter over the former for phenotype prediction with large-scale dataset ($n \approx 10$k).

These first conclusions were crucial for developing our main paradigm based on transfer learning. In Chap. 3, we hypothesized that DNN could learn a transferable representation from a large-scale dataset of healthy controls (now easily available), to discriminate patients from controls in a second fine-tuning phase. From this point-of-view, patients with brain disorder are viewed as deviation from a manifold formed by the healthy population, following a dimensional approach (*i.e.,* assuming a continuous spectrum across brain disorders, potentially sharing common dimensions in the latent space). From this idea, we have developed new tools for unsupervised representation learning of brain images based on contrastive learning (CL). This discriminative approach does not require pixel-level generation (like generative models do) but rather the definition of positive and negative distributions in order to explicitly determine

semantically similar (positive) samples and dissimilar (negative) samples. In our context, we proposed a weakly-supervised approach by including auxiliary non-imaging information (such as phenotype) in the definition of positive and negative distributions. Imaging samples with similar auxiliary variables are mapped closely in the representation space, assuming they share more anatomical traits than two samples with very dissimilar auxiliary information. From an Information Bottleneck point-of-view, we compress input data to maximize the information shared between image representation and its auxiliary variable. We found that the resulting pre-trained model was versatile, producing state-of-the-art performance for discriminating patients and controls for three psychiatric disorders (schizophrenia, bipolar disorder, ASD) and Alzheimer's disease.

In the second part of Chap. 3, we have continued our exploration of contrastive models by giving first theoretical guarantees for generalization performance on new supervised tasks. Notably, we demonstrated tight bounds between unsupervised and supervised objectives under strong assumptions on data augmentation used in the original CL framework to define positive distribution. Then, we demonstrated that we can relax this hypothesis if we introduce prior information (e.g. given as auxiliary variable or by a generative model) that relates intra-class samples through a kernel function. This theory bridges the gap between generative models and CL for learning representations. We empirically show the validity of this theory on several benchmarks with natural and brain images. It is a first (modest) step towards a better understanding of CL models.

Based on our in-depth analysis of DNN in Chap. 2, we have built OpenBHB, a new large-scale brain imaging dataset designed for supervised representation learning with site-effect removal. We present its unique properties in Chap. 4 notably in terms of size, heterogeneity (multi-site/multi-location images), and pre-processing. In Chap. 2 we have shown poor cross-site generalization for all machine learning models, potentially due to a high over-fitting effect on acquisition settings/scanner brain images are coming from. Consequently, we hope this large dataset offers a high quality benchmark resource for the neuroimaging community to tackle brain age prediction and site debiasing while improving reproducibility of new ML models. Along with OpenBHB, we proposed and setup a permanent challenge presented in the NeuroImage special edition "*Benchmarks for Machine Learning in Neuroimaging*". This challenge offers a unique way to rank and compare submitted models on the same imaging resource, for both brain age prediction and site debiasing through a public leaderboard and novel metrics based on representation learning tools (e.g. linear probing).

Perspectives

4.5.3 Integrating phenotype/genotype knowledge for learning representations

We showed in Chap. 3 that contrastive model can greatly benefit from auxiliary information (e.g. age) to learn deep representations from brain imaging on the healthy population. We showed that these phenotyping information can be accurately decoded from imaging features using our framework and it can improve the transfer capacity on clinical datasets. An interesting question is whether integration of additional phenotype/genotype variables (e.g. cognition, life style, education level, etc.) would also improve deep representations emerging from brain imaging to discriminate patients and controls at subject-level. If it is the case, a second subsequestion question would be which non-imaging variables influence the most the final representation for a given downstream task. The current framework allows to answer such question. Indeed, the kernel function measuring the similarity between auxiliary variables can be adapted to the multi-dimensional case (*e.g.,* with a product of 1D Gaussian kernels applied on each auxiliary information). By pre-training the model on different sets of phenotype/genotype, we can evaluate the different representations that emerge (*e.g.,* with linear probing on several brain disorders) and check whether these representations improve brain pathology decoding or not. Nevertheless, testing all combinations of auxiliary variables is computationally expansive and it may be too costly. Another line of research would be to learn directly the kernel during pre-training. For instance, a parametric family of kernels can be set (such as Gaussian family with variance as parameter) and the parameters can be learned with gradient-descent. After pre-training, the kernel can be further analyzed and related to the quality of the model's representation.

4.5.4 Contrastive learning with multi-modal brain imaging

In Chap. 3, we demonstrated that contrastive learning is well-suited for pre-training deep models on large-scale brain MRI dataset. While we focused our analysis on anatomical imaging, other modalities such as functional and diffusion MRI provide additional features that would allow i) new neuromarkers discovery specific to mental illnesses inside brain networks (such as default-mode network, central executive network and salience network for schizophrenia [267]); ii) decoupling environmental from genetic variability in neuro-developmental disorders using both brain folding patterns extracted from sMRI (assuming that it integrates mostly genetic variability [34]) and structural connectivity from dMRI (integrating environmental and genetic variability). A natural approach is to consider each modality as a view of the same instance in the contrastive framework. This model extracts joint information between several modalities of the same instance. Nonetheless, finding modality-specific information remains a challenge that needs to be addressed. First work focusing on Alzheimer's disease [94] started to emerge but it did not tackle this critical issue. A first idea would be to both i) align inter-modality

representation and ii) intra-modality representation with CL objectives and to fuse all representations to a common space. This idea has been applied very recently to vision-language model [308], yielding good performance at image-text retrieval and visual question-answering. In our context, brain region patterns could be discovered for brain disorders with such multi-modal approach, for instance by relating cortical and sub-cortical neuroanatomical signatures with white matter disconnections between regions using dMRI and sMRI of the same subject.

4.5.5 Debiasing deep representations

During our analysis of deep neural networks in Chap. 2, we observed poor cross-site generalization performance and high over-fitting on acquisition settings/scanner for all prediction tasks. This recurrent issue in multi-site studies led us to create the OpenBHB challenge (described Chap. 4) but current solutions remain unsatisfactory. A lot of recent works using deep models are emerging both in neuroimaging [77] and computer vision [273] (e.g. with applications to fairness and trustworthy AI) but there is currently little or no consensus in these two fields. We proposed a first solution based on contrastive learning by adding a regularization term during optimization that strongly constrains bias-conflicting samples with the same class attribute to be aligned in the latent space. Nonetheless, while supervised contrastive learning is well-suited for classification tasks, its derivation for regression problems is not obvious since the definition of positive samples is not clear. Our first solution consisted in defining the positive distribution with a kernel (like we did in Chapter 3 with auxiliary variables). This way, subjects with close age should be closer in the latent space and we can re-define the distance between positives using this kernel. Other solutions could be imagined to tackle contrastive learning for regression (such as [299]).

More broadly, our approach belonged to regularization-based methods but other approaches can be envisioned for debiasing. In particular, adversarial training (*e.g.,* BlindEye [7]), where a classifier is trained in an adversarial manner on top of the representation to predict the bias (acquisition site here), may provide a generic solution for debiasing duing optimization. Such adversarial technique has been crafted recently for neuroimaging data [82] however it still requires three different stages during optimization, which can lead to high instability during training (corroborated by the large variance observed in the reported results). Another interesting direction is by exploiting training dynamics, starting from a simple observation on synthetic experiments: the bias is generally easy to capture and it is learned first during training. The idea is thus to learn debiased representation from a biased one (*e.g.,* [205]) by putting more weights on biased-conflicting samples (similarly with what we proposed but with a re-weighting scheme instead of a regularization technique).

All these techniques provide interesting research directions for brain image analysis when working at a large-scale with pooled multi-site datasets. In the end, they allow learning a debiased representation of neuroimaging data and to apply this model on new data, potentially

acquired on new sites, unseen during training. Such models are crucial in a real clinical scenario where we cannot reasonably assume to have the same scanners and acquisition settings in all hospitals as the ones used to produced our training data.

4.5.6 Future works for learning representations with AI: from continuous to symbolic approach

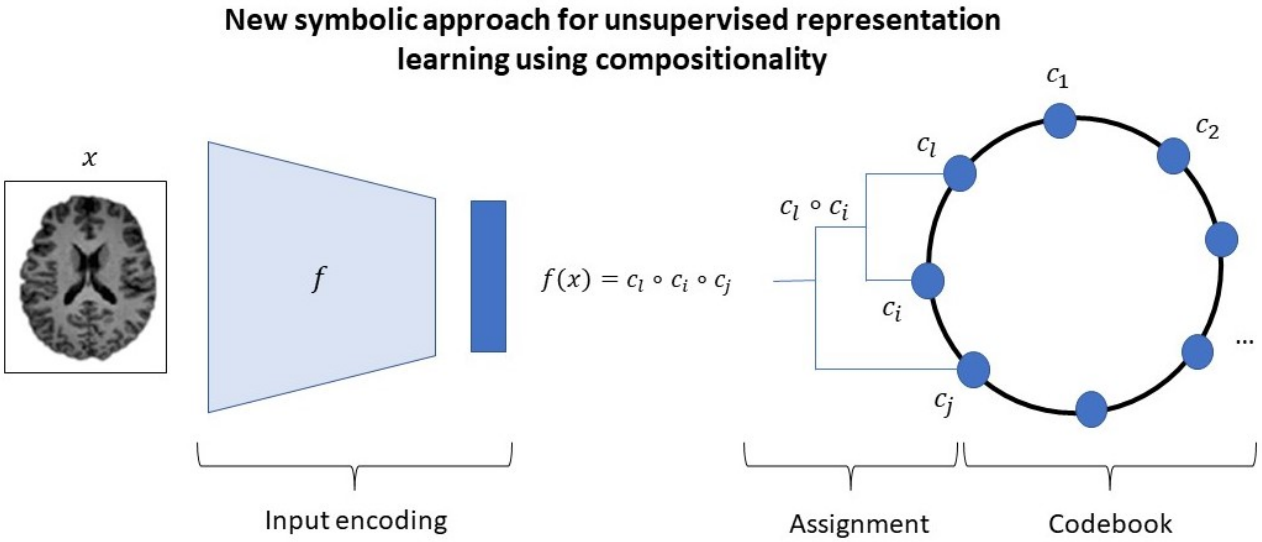


Figure 4.11: Overview of the proposed framework for learning representations using compositionality in the latent space based on a discrete set of symbols. We assume the latent space to be low-dimensional embedding on a hyper-sphere. The composition operator $\circ$ defines the composition structure of input representation $f(x)$. The symbols $\{c_k\}_{k=1}^{K}$ are vectors, either fixed at initialisation or learned during training but always uniformly spread over the hyper-sphere. Training consists in imposing invariance of $f(x)$ over a set of transformations $\mathcal{T}$ (such as aggressive crop or cutout for images), *i.e.*, $\forall t \sim \mathcal{T}, f(t(x)) = f(x)$. Assignment between $f(x)$ and the subset of symbols $\{c_k\}_{k \in I(x)}$ needs to be specified according to the choice of $\circ$. For simple addition $+$ between vectors, optimal transport algorithm can be used like in SwAV [43].

Unsupervised representation learning paradigm for visual representation is currently based on image invariance to transformations (learning from repetition with variation). All recent "instance-based" approaches (SimCLR [52], MoCo [136], BYOL [120], etc.) impose similarity between internal representations of aggressively transformed input to learn without supervised signal. These approaches assume that each image has its own representation, distinct from any other images. From this perspective, infinite number of images would lead to infinite number of representations. On the other hand, cluster-based/prototypical approaches like SwAV [43] assume that **it exists a *finite* number of internal codes such that any input can be represented by a combination of these codes.** Moreover, variations (i.e. transformations) of the same input produce the same combination of codes. Interestingly, in practice, the codebook size required to reach state-of-the-art results on vision benchmarks with SwAV is very small compared to the dataset size (typically 3-4K vs 1M for ImageNet [74]). Such model is a first step towards linking continuous high-dimensional input to small discretized latent codes

that compress the relevant semantic information in the same way we, as humans, might do in our day-to-day life. Indeed, current theory [73] proposed by S. Dehaene in cognitive science states: *we propose that humans are characterized by a specific ability to attach discrete symbols to mental representations and to combine those symbols into nested recursive structures called mental programs, the compositional rules of which define a language of thought. Humans develop multiple such languages of thought in various domains (linguistic, musical, mathematical...).*

This view, also largely developed by Chomsky with generative grammar for language [59], is in line with a symbolic approach for representation learning using deep neural networks. We argue, in line with P. Smolensky [260], that combining continuous computations with discrete representations using a symbolic approach will lead to the next generation of AI system, in particular for neuroimaging. It would notably allow:

1. a better integration of the compositionality principle, which may reduce drastically the number of training examples required to learn representations (a bottleneck in particular for clinical applications)

2. better interpretability of deep models, since input representation would rely on a finite number of symbols that could be individually investigated (e.g. to link brain networks with symbols)

More concretely, with the previous notion of codebook for learning representation, a representation $f(x)$ of an input x (such as brain image) could be decomposed over a *finite* number of symbols (a.k.a. codes) $\{c_k\}_{k=1}^K$ that could be fixed or learned. In SwAV, the learning procedure consists in i) decomposing linearly $f(x)$ over $\{c_k\}$; ii) preserving uniformity between these symbols and iii) imposing representation invariance over input transformations. In this approach, the weights associated to each symbol can be viewed as a probability of belonging to this symbol (like in soft clustering). However, we argue that there is no notion of compositionality or recurvise nested structure for decomposing $f(x)$ over $\{c_k\}$. To specifically define the "language of thought" in such representation, a grammar is needed–that is, the compositional rules over these symbols that allows to perform mental program on a given task. One idea is to decompose the representation $f(x)$ using a binary tree structure whose leafs are a subset of symbols $\{c_k\}$ and the nodes is the result of a composition operator $\circ$ between two intermediate representations. The learning algorithm would consist, as before, in imposing (1) invariance of representation $f(x)$ over a set of transformations while (2) preserving uniformity between $\{c_k\}$. In particular, this paradigm implies that, after training, each binary tree associated to the representation of an input is invariant to a group of transformations. Considering the current performance of SwAV for unsupervised learning of visual representation, we argue that constraints (1) and (2) are sufficient to obtain good representations.

The composition operator $\circ$ ultimately sets how $f(x)$ should be decomposed over $\{c_k\}$. For instance, if $\circ$ is the standard addition operator between vectors, then $f(x)$ is a weighted sum over a subset of symbols with only discrete weights. These weights could be learned using optimal

transport (like in SwAV) by finding the optimal transportation polytope from representation $f(x)$ to symbols, restrained to only integer weights (*e.g.,* using a rounding system). However, multiple binary trees (over the same symbols) could lead to the same representation $f(x)$ so there would not be identifiability using this composition operator. Other composition operators have been proposed in cognitive science (*e.g.,* Mitchell & Lapata proposed additive models [202] such as $c_i \circ c_j = Ac_i + Bc_j$) and each one implies a specific optimal transport algorithm to map $f(x)$ to symbols $\{c_k\}$. The question becomes, which composition operator is best suited for learning (visual) representations, *i.e.,* which operator $\circ$ (and subsequent optimal transport algorithm) leads to best generalization performance? Can we improve current SwAV model by finding such composition operator? Some answers to these questions may open the door to new AI solutions that integrate both compositionality principle and invariance to a group of transformations at its core, and it may lead to a new generation of AI systems.

Appendix A

First Appendix

A.1 Bayesian Inference

Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of labelled examples, we define a posterior distribution $p(y|x, \mathcal{D})$ for a new input x that can be computed with a neural network $f_\theta(x)$ with a prior distribution $p(\theta)$ on its parameters. To compute this posterior distribution, we can use Bayesian Inference. Here, we prove that the following equality holds for an input x:

$$p(y|x, \mathcal{D}) = \int p(y|x, \theta)p(\theta|\mathcal{D})d\theta \tag{A.1}$$

PROOF. The proof essentially comes from the Bayesian equality $p(y, \theta|Z) = \frac{p(y,\theta,Z)}{p(Z)} \cdot \frac{p(\theta,Z)}{p(\theta,Z)} = p(y|\theta, Z)p(\theta|Z)$. By setting $Z = (x, D)$, we have:

$$\begin{aligned} p(y|x, \mathcal{D}) &= \int p(y, \theta|x, \mathcal{D})d_\theta \\ &= \int p(y|x, \mathcal{D}, \theta)p(\theta|x, \mathcal{D})d_\theta \\ &= \int p(y|x, \theta)p(\theta|\mathcal{D})d_\theta \end{aligned}$$

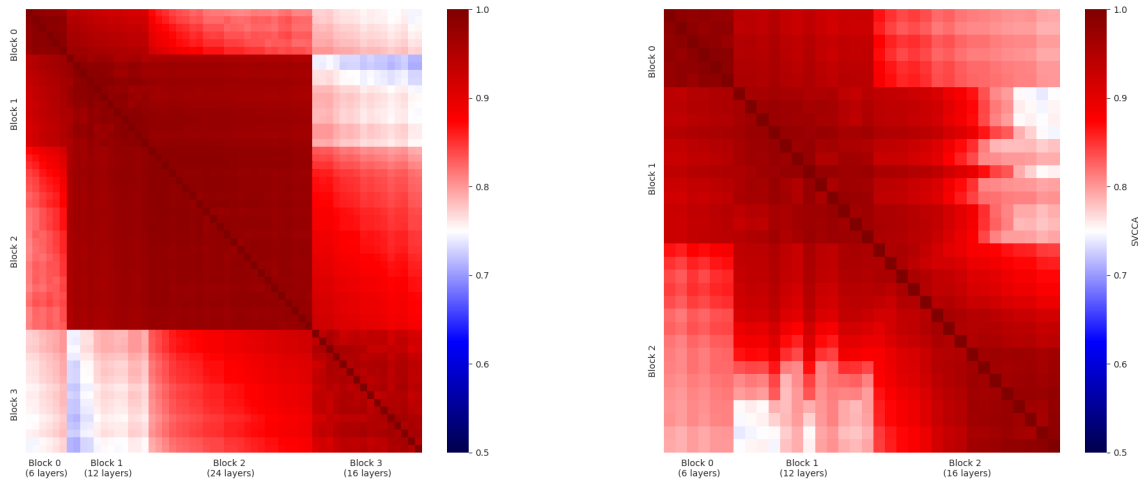
Last equality holds since $p(y|x, \mathcal{D}, \theta)$ does not depend on $\mathcal{D}$ once the parameters θ are fixed and $p(\theta|x, \mathcal{D})$ does not depend on input x when training samples $\mathcal{D}$ are provided.

A.2 Introduction of tiny-DenseNet

Analysis of DenseNet121: as we wanted to give a tiny version of DenseNet (121 layers and 11M parameters), we analyzed its internal representation on Dx problem. In order to analyze the representation learnt inside this network, we computed the Singular Vector Canonical Correlation Analysis (SVCCA) [229] between the outputs of all pairs of layer inside every block. Formally, we define a set of neurons $\{\mathbf{z}_i^l\}_{i\in[1..hwcd]}$ for each layer l where (c, h, w, d) represent the

number of channels, height, width and depth of the feature maps of layer l respectively; and $\mathbf{z}_i^l = (\mathbf{z}_i^l(x_1), ..., \mathbf{z}_i^l(x_N)) \in \mathbb{R}^N$ is the response of neuron i to the entire test set (of size N). In this way, we can compute the CCA between 2 blocks of data $\{\mathbf{z}_i^{l_1}\}_{i\in[1..h_1w_1c_1d_1]}$ and $\{\mathbf{z}_i^{l_2}\}_{i\in[1..h_2w_2c_2d_2]}$ for 2 layers l_1 and l_2 since all vectors lie in the same space $\mathbb{R}^N$ (we also computed a Singular Value Decomposition (SVD) before the computation of the CCA to remove the noisy neurons, as described in [229]). We chose to keep only 50% of the explained variance since $N \ll hwcd$ in our experiments ($N = 394$ and $hwcd > 10^4$) and we observed that a lot of neurons were noisy. Results are plotted in figure A.1a.

Tiny-DenseNet: we first observed that the blocks 1 and 2 (starting from 0) of DenseNet121 were highly correlated, which suggested a redundancy. In particular, it suggested that the features learnt inside the 3$^{\text{rd}}$ block were just copied from the second block and the specialization of the network to the prediction task did not occur in block 2. It was then natural to remove the block 2 from DenseNet121, assuming that the receptive field of a neuron before the FC layer would remain big enough for the 3 clinical tasks (its size is $32 \times 32 \times 32$ for a an input size $128 \times 128 \times 128$ with DenseNet121 and it is halved when we remove the 3$^{\text{rd}}$ block). Also, we halved the growth rate from $k = 32$ to $k = 16$ and we called the resulting network *tiny-DenseNet*, as it is 10× smaller than DenseNet. As before, we plotted the SVCCA between the internal layer outputs of tiny-DenseNet in figure A.1b and we noticed that, differently from DenseNet121 in figure A.1a, the strong correlation between blocks disappeared.



(a) DenseNet121

(b) tiny-DenseNet

Figure A.1: Internal representation of DenseNet and its tiny version. The SVCCA is computed between each pair of layers. Networks are trained on Dx.

Appendix B

Second Appendix

B.1 Inequality between InfoNCE and NCE loss

Property 1. *InfoNCE loss $\mathcal{L}_{InfoNCE}$ upper bounds the NCE loss $\mathcal{L}_{NCE}$ such that $\mathcal{L}_{NCE} \leq \mathcal{L}_{InfoNCE} + \log(1+e) + O\left(\frac{1}{N}\right)$*

PROOF.

We recall the definition of the two losses, for N pairs of views $(v_1^i, v_2^i) \sim p(V_1, V_2)$:

$$\mathcal{L}_{NCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log h_\theta(v_1^i, v_2^i) + \frac{1}{N-1}\sum_{j\neq i}\log(1 - h_\theta(v_1^i, v_2^j))\right) \tag{B.1}$$

$$\mathcal{L}_{InfoNCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{-E_\theta(v_1^i, v_2^i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{-E_\theta(v_1^i, v_2^j)}} \tag{B.2}$$

with $h_\theta(\cdot,\cdot) = \sigma(E_\theta(\cdot,\cdot))$ and σ is the Sigmoid function. We first start by re-writing NCE loss as:

$$\begin{aligned}
\mathcal{L}_{NCE} &= -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{1}{1+e^{E_\theta(v_1^i, v_2^i)}} + \frac{1}{N-1}\sum_{j\neq i}\log\frac{e^{E_\theta(v_1^i, v_2^j)}}{1+e^{E_\theta(v_1^i, v_2^j)}}\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{e^{-E_\theta(v_1^i, v_2^i)}}{e^{-E_\theta(v_1^i, v_2^i)}+1} + \frac{1}{N-1}\sum_{j\neq i}\log\frac{1}{e^{-E_\theta(v_1^i, v_2^j)}+1}\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{e^{-E_\theta(v_1^i, v_2^i)}}{e^{-E_\theta(v_1^i, v_2^i)}+1} + \frac{1}{N-1}\sum_{j\neq i}\log\frac{1}{e^{-E_\theta(v_1^i, v_2^j)}+1}\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(-E_\theta(v_1^i, v_2^i) - \frac{1}{N-1}\sum_{j=1}^{N}\log\left(e^{-E_\theta(v_1^i, v_2^j)}+1\right)\right)
\end{aligned} \tag{B.3}$$

We can then re-write the denominator of InfoNCE loss and apply Jensen's inequality:

$$\begin{aligned}
\log \frac{1}{N}\sum_{j=1}^{N} e^{-E_\theta(v_1^i, v_2^j)} &= \log\left(\frac{1}{N}\sum_{j=1}^{N}\left(e^{-E_\theta(v_1^i, v_2^j)}+1\right)-1\right) \\
&\overset{(1)}{\geq} \log\left(\frac{1}{N}\sum_{j=1}^{N}\left(e^{-E_\theta(v_1^i, v_2^j)}+1\right)\right)-K \\
&\overset{(2)}{\geq} \frac{1}{N}\sum_{j=1}^{N}\log\left(e^{-E_\theta(v_1^i, v_2^j)}+1\right)-K \\
&\geq \frac{1}{N-1}\sum_{j=1}^{N}\log\left(e^{-E_\theta(v_1^i, v_2^j)}+1\right)-K+O\left(\frac{1}{N}\right) \qquad \text{(B.4)}
\end{aligned}$$

Where (1) stands since $\log(x-1) \geq \log(x) - K, \forall x \geq 1+e^{-1}$ with $K = \log(1+e)$ and (2) is by Jensen's inequality applied to convex function $-\log$. By combining eq. B.3 and B.4, we have:

$$\begin{aligned}
\mathcal{L}_{NCE} &\leq -\frac{1}{N}\sum_{i=1}^{N}\left(e^{-E_\theta(v_1^i, v_2^i)} - \log\sum_{j=1}^{N}\frac{1}{N}e^{-E_\theta(v_1^i, v_2^j)}\right)+K+O\left(\frac{1}{N}\right) \\
&= \mathcal{L}_{InfoNCE}+K+O\left(\frac{1}{N}\right)
\end{aligned}$$

B.2 Equivalence between y-Aware InfoNCE and SupCon in discrete case

Theorem 7. *Let $(v_1^i, v_2^i, y_i)_{i\in[1..N]} \overset{\text{iid}}{\sim} p(V_1, V_2, Y)$ with $Y \in \{1..K\}$ a discrete auxiliary variable ($K \in \mathbb{N}$). Then y-Aware InfoNCE loss is the SupCon loss [170] and it is a negative estimator of the mutual information $I(V_1, V_2)$ under the conditional independence assumption $p(V_1, V_2|Y) = p(V_1|Y)p(V_2|Y)$:*

$$I^y_{NCE}(V_1, V_2) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_{y_i}|^2}\sum_{k_1,k_2\in C_{y_i}}\log\frac{e^{f_\theta(v_1^{k_1})\cdot f_\theta(v_2^{k_2})}}{\frac{1}{N}\sum_{j=1}^{N}e^{f_\theta(v_1^{k_1})\cdot f_\theta(v_2^{j})}} \qquad \text{(B.5)}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_{y_i}|}\sum_{j\in C_{y_i}}\log\frac{e^{f_\theta(v_1^{i})\cdot f_\theta(v_2^{j})}}{\frac{1}{N}\sum_{k=1}^{N}e^{f_\theta(v_1^{i})\cdot f_\theta(v_2^{k})}} = -\mathcal{L}_{SupCon} \qquad \text{(B.6)}$$

Where $C_y = \{i \in [1..N] | y_i = y\}$.

PROOF. To prove this equality, we separate the first sum according to the label of each sample

$i \in [1..N]$. Let $F_\theta(i,j) = \log \frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^j)}}{\frac{1}{N}\sum_{k=1}^N e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}}$, then we have:

$$\begin{aligned}
I^y_{NCE}(V_1,V_2) &= \frac{1}{N}\sum_{i=1}^N \frac{1}{|C_{y_i}|^2}\sum_{k_1,k_2\in C_{y_i}} F_\theta(k_1,k_2) \\
&= \frac{1}{N}\sum_{y=1}^K \sum_{i\in C_y}\frac{1}{|C_{y_i}|^2}\sum_{k,j\in C_{y_i}} F_\theta(k,j) \\
&\overset{(1)}{=} \frac{1}{N}\sum_{y=1}^K \sum_{i\in C_y}\sum_{k,j\in C_y}\frac{1}{|C_y|^2} F_\theta(k,j) \\
&= \frac{1}{N}\sum_{y=1}^K |C_y| \sum_{k,j\in C_y}\frac{1}{|C_y|^2} F_\theta(k,j) \\
&= \frac{1}{N}\sum_{y=1}^K \sum_{k\in C_y}\sum_{j\in C_y}\frac{1}{|C_y|} F_\theta(k,j) \\
&= \frac{1}{N}\sum_{k=1}^N \frac{1}{|C_{y_k}|}\sum_{j\in C_{y_k}} F_\theta(k,j) = -\mathcal{L}_{SupCon}
\end{aligned}$$

Where (1) stands because $C_y = C_{y_i}$ for $i \in C_y$ by definition. This relates our y-Aware InfoNCE estimator to SupCon. Since $I^y_{NCE}(V_1,V_2)$ is an estimator of the mutual information $I(V_1,V_2)$ under conditional independence assumption, so is SupCon.

B.3 Contrastive Learning optimizes alignment and uniformity

Theorem 8. *InfoNCE converges, as the number of pairs N increases, to:*

$$\begin{aligned}
\mathcal{L}_{InfoNCE} &= -\mathbb{E}_{(v_1^i,v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)}\left(\frac{1}{N}\sum_{i=1}^N \log\frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^i)}}{\frac{1}{N}\sum_{k=1}^N e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}}\right) \\
&\xrightarrow[N\to\infty]{} -\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)}\left[f_\theta(v_1)\cdot f_\theta(v_2)\right] + \mathbb{E}_{v_1\sim p(V_1)}\log \mathbb{E}_{v_2\sim p(V_2)} e^{f_\theta(v_1)\cdot f_\theta(v_2)}
\end{aligned}$$

PROOF. To prove this theorem, we first split $\mathcal{L}_{InfoNCE}$ into 2 terms and we then use Strong Law of Large Numbers (SLLN):

$$\mathcal{L}_{InfoNCE} = -\mathbb{E}_{(v_1^i, v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)} \left(\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{f_\theta(v_1^i)\cdot f_\theta(v_2^i)}}{\frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}} \right)$$

$$= -\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)}(f_\theta(v_1)\cdot f_\theta(v_2)) + \mathbb{E}_{(v_1^i, v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)} \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}$$

$$= -\mathbb{E}_{(v_1,v_2)\sim p(V_1,V_2)}(f_\theta(v_1)\cdot f_\theta(v_2)) +$$

$$\mathbb{E}_{(v_1^i)_{i\in[1..N]}\sim p(V_1)} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{v_2^i,(v_2^j)_{j\neq i}\sim p(V_2|V_1)p(V_2)^{N-1}} \log \frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)}$$

Let $i \in [1..N]$ and $v_1^i \sim p(V_1)$. For $N-1$ samples $(v_2^j)_{j\neq i} \sim p(V_2)$ and $v_2^i \sim p(V_2|V_1)$, we remark that $\lim_{N\to\infty} \log \frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)} = \log \mathbb{E}_{v_2\sim p(V_2)} e^{f_\theta(v_1^i)\cdot f_\theta(v_2)}$ by SLLN and continuous mapping theorem. Then, we obtain:

$$\lim_{N\to\infty} \mathbb{E}_{(v_1^i, v_2^i)_{i\in[1..N]}\sim p(V_1,V_2)} \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{N}\sum_{k=1}^{N} e^{f_\theta(v_1^i)\cdot f_\theta(v_2^k)} = \lim_{N\to\infty} \mathbb{E}_{(v_1^i)_{i\in[1..N]}\sim p(V_1)} \frac{1}{N}\sum_{i=1}^{N} \log \mathbb{E}_{v_2\sim p(V_2)} e^{f_\theta(v_1^i)\cdot f_\theta(v_2)}$$

$$= \mathbb{E}_{v_1\sim p(V_1)} \log \mathbb{E}_{v_2\sim p(V_2)} e^{f_\theta(v_1)\cdot f_\theta(v_2)}$$

Which concludes the proof by taking the previous decomposition.

B.4 More Empirical Evidence with Decoupled Uniformity objective

In this section, we provide additional empirical evidence to confirm several claims and arguments developed in the main text.

B.4.1 Multi-view Contrastive Learning with Decoupled Uniformity

When the intra-class connectivity hypothesis is full-filled, we showed that Decoupled Uniformity loss can tightly bound the classification risk for well-aligned encoders (see Theorem 4). Under that hypothesis, we consider the standard empirical estimator of $\mu_x \approx \sum_{v=1}^{V} f(x^{(v)})$ for V views. Using all SimCLR augmentations, we empirically verify that increasing V allows for: 1) a better estimate of μ_x which implies a faster convergence and 2) better SOTA results on both small-scale (CIFAR10, CIFAR100, STL10) and large-scale (ImageNet100) vision datasets. We always use batch size $n = 256$ for all approaches with ResNet18 backbone for CIFAR10, CIFAR100 and STL10 and ResNet50 for ImageNet100. We report the results in Table B.1.

B.4.2 Kernel choice on RandBits experiment

In our experiments on RandBits, we used RBF Kernel in Decoupled Uniformity but other kernels can be considered. Here, we have compared our approach with a cosine kernel on Randbits with $k = 10$ and $k = 20$ bits. There is no hyper-parameter to tune with cosine. From

Model	CIFAR-10		CIFAR-100		ImageNet100		STL10	
	$e = 200$	$e = 400$	$e = 200$	$e = 400$	$e = 200$	$e = 400$	$e = 200$	$e = 400$
SimCLR[52]	79.4	81.75	48.89	53.02	65.30	66.52	76.99	79.02
BYOL[120]	80.14	81.97	51.57	53.65	72.20	72.26	77.62	79.61
Decoupled Unif (2 views)	82.43	**85.82**	54.01	58.89	71.98	72.24	78.12	79.89
Decoupled Unif (4 views)	84.99	85.34	57.23	59.07	72.08	75.00	78.25	**80.47**
Decoupled Unif (8 views)	**86.50**	85.80	**59.63**	**59.74**	**74.70**	**75.00**	**79.82**	80.30

Table B.1: A better approximation of centroids μ_x (i.e. increasing number of views) when augmentation overlap hypothesis is (nearly) full-filled implies faster convergence. All models are pre-trained with batch size $n = 256$. We use ResNet18 backbone for CIFAR10, CIFAR100, STL10 and ResNet50 for ImageNet100. We report linear evaluation accuracy (%) for a given number of epochs e.

Table B.2, we see that cosine gives comparable results for $k = 10$ bits with RBF but it is not appropriate for $k = 20$ bits.

Kernel	10 bits	20 bits
RBFKernel($\sigma = 1$)	$66.25_{\pm 0.17}$	$9.91_{\pm 0.13}$
RBFKernel($\sigma = 30$)	$67.21_{\pm 0.29}$	$66.46_{\pm 0.19}$
RBFKernel($\sigma = 50$)	$68.42_{\pm 0.51}$	$68.58_{\pm 0.17}$
CosineKernel	$66.56_{\pm 0.45}$	$9.68_{\pm 0.18}$

Table B.2: Linear evaluation after training on RandBits-CIFAR10 with ResNet18 for 200 epochs. RBF and Cosine kernels are evaluated.

B.5 Geometrical Considerations about Decoupled Uniformity

In this section, we provide a geometrical understanding of Decoupled Uniformity loss from a metric learning point of view. In particular, we consider the Log-Sum-Exp (LSE) operator often used in CL as an approximation of the maximum.

We consider the finite-samples case with N original samples $(x_i)_{i\in[1..N]} \overset{\text{i.i.d.}}{\sim} p(X)$ and L views $(v_i^{(l)})_{l\in[1..L]} \overset{\text{i.i.d.}}{\sim} p_{\mathcal{A}}(V|x_i)$ for each sample x_i. We make an abuse of notations and set $\mu_i = \frac{1}{L}\sum_{l=1}^{L} f(v_i^{(l)})$. Then we have:

$$\begin{aligned}\hat{\mathcal{L}}^{d}_{unif} &= \log \frac{1}{N(N-1)} \sum_{i\neq j} \exp\left(-||\mu_i - \mu_j||^2\right) \\ &= \log \frac{1}{N(N-1)} \sum_{i\neq j} \exp\left(-s_i^+ - s_j^+ + 2s_{ij}^-\right)\end{aligned} \tag{B.7}$$

where $s_i^+ = ||\mu_i||^2 = \frac{1}{L^2}\sum_{l,l'} s(v_i^{(l)}, v_i^{(l')})$, $s_{ij}^- = \frac{1}{L^2}\sum_{l,l'} s(v_i^{(l)}, v_j^{(l')})$ and $s(\cdot,\cdot) = \langle f(\cdot), f(\cdot)\rangle_2$ is viewed as a similarity measure.

From a metric learning point-of-view, we shall see that minimizing Eq. B.7 is (almost) equivalent to looking for an encoder f such that the sum of similarities of all views from the same anchor (s_i^+ and s_j^+) are higher than the sum of similarities between views from different

instances s_{ij}^-:

$$s_i^+ + s_j^+ > 2s_{ij}^- + \epsilon \quad \forall i \neq j \tag{B.8}$$

where ϵ is a margin that we suppose "very big" (see hereafter). Indeed, this inequality is equivalent to $-\epsilon > 2s_{ij}^- - s_i^+ - s_j^+$ for all $i \neq j$, which can be written as :

$$\arg\min_f \max(-\epsilon, \{2s_{ij}^- - s_i^+ - s_j^+\}_{i,j\in[1..N],j\neq i})$$

This can be transformed into an optimization problem using the LSE (log-sum-exp) approximation of the max operator:

$$\arg\min_f \log\left(\exp(-\epsilon) + \sum_{i\neq j} \exp\left(-s_i^+ - s_j^+ + 2s_{ij}^-\right)\right)$$

Thus, if we use an infinite margin ($\lim_{\epsilon\to\infty}$) we retrieve exactly our optimization problem with Decoupled Uniformity in Eq.B.7 (up to an additional constant depending on N).

B.6 Additional general guarantees on downstream classification

B.6.1 Optimal configuration of supervised loss

In order to derive guarantees on a downstream classification task $\mathcal{D}$ when optimizing our unsupervised decoupled uniformity loss, we define a supervised loss that measures the risk on a downstream supervised task. We prove in the next section that the minimizers of this loss have the same geometry as the ones minimizing cross-entropy and SupCon [170]: a regular simplex on the hyper-sphere [118]. More formally, we have:

Lemma 9. *Let a downstream task $\mathcal{D}$ with C classes. We assume that $C \leq d+1$ (i.e., a big enough representation space), that all classes are balanced and the realizability of an encoder $f^* = \arg\min_{f\in\mathcal{F}} \mathcal{L}_{sup}(f)$ with $\mathcal{L}_{sup}(f) = \log \mathbb{E}_{y,y'\sim p(Y)p(Y')} e^{-||\mu_y - \mu_{y'}||^2}$, and $\mu_y = \mathbb{E}_{x\sim p(X|Y=y)}\mu_x$. Then the optimal centroids $(\mu_y^*)_{y\in\mathcal{Y}}$ associated to f^* make a regular simplex on the hypersphere $\mathbb{S}^{d-1}$ and they are perfectly linearly separable, i.e $\min_{(w_y)_{y\in\mathcal{Y}}\in\mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathbb{1}(w_y \cdot \mu_y^* < 0) = 0$. Proof in the next section.*

This property notably implies that we can realize 100% accuracy at optima with linear evaluation (taking the linear classifier $g(x) = W^* f^*(x)$ with $W^* = (\mu_y^*)_{y\in\mathcal{Y}} \in \mathbb{R}^{C\times d}$).

B.6.2 General guarantees of Decoupled Uniformity

In its most general formulation, we tightly bound the previous supervised loss by Decoupled Uniformity loss $\mathcal{L}_{unif}^d$ depending on a variance term of the centroids μ_x conditionally to the labels:

Theorem 10. *(Guarantees for a given downstream task) For any $f \in \mathcal{F}$ and augmentation $\mathcal{A}$ we have:*

$$\mathcal{L}^d_{unif}(f) \leq \mathcal{L}_{sup}(f) \leq 2\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y) + \mathcal{L}^d_{unif}(f) \leq 4\mathbb{E}_{x,x'\sim p(X|y)p(X'|y)}||\mu_x - \mu_{x'}|| + \mathcal{L}^d_{unif}(f) \tag{B.9}$$

where $\mathrm{Var}(\mu_x^j|y) = \mathbb{E}_{x\sim p(X|y)}(\mu_x^j - \mathbb{E}_{x'\sim p(X|y)}\mu_{x'}^j)^2$, $y = \arg\max_{y'\in\mathcal{Y}} \mathrm{Var}(\mu_x^j|y')$ *and* μ_x^j *is the* j*-th component of* $\mu_x = \mathbb{E}_{v\sim p_{\mathcal{A}}(V|x)} f(v)$*. Proof in the next section.*

Intuitively, it means that we will achieve good accuracy if all centroids $(\mu_x)_{x\in\mathcal{X}}$ for samples $x \in \mathcal{X}$ in the same class are not too far. This theorem is very general since we do not require the intra-class connectivity assumption on $\mathcal{A}$; so any $\mathcal{A} \subset \mathcal{A}^*$ can be used.

B.7 Experimental Details

We provide a detailed pseudo-code of our algorithm as well as all experimental details to reproduce the experiments run in the manuscript.

B.7.1 Pseudo-code

Algorithm 1 Pseudo-code of the algorithm

Require: Batch of images $(x_1, ..., x_N) \in \mathcal{X}$, augmentation module $\mathcal{A}$

$K_N \leftarrow (K(x_i, x_j))_{i,j\in[1..N]}$ ▷ Compute the kernel matrix

$\alpha \leftarrow (K_N + N\lambda \mathbf{I}_N)^{-1} K_N$ ▷ Compute weights for centroid estimation

$v_i^{(1)}, ..., v_i^{(L)} \overset{\text{i.i.d.}}{\sim} p_{\mathcal{A}}(V|x_i)$ ▷ Sample L views per image

$F \leftarrow (\frac{1}{L}\sum_{l=1}^{L} f(v_i^{(l)}))_{i\in[1..N]}$ ▷ Compute the averaged image representation

$\hat{\mu} \leftarrow \alpha F$ ▷ Centroid estimation

$\hat{\mathcal{L}}^d_{unif} \leftarrow \log \frac{1}{N(N-1)} \sum_{i\neq j} \exp(-||\hat{\mu}_i - \hat{\mu}_j||^2)$ ▷ Kernel Decoupled Uniformity loss

return $\hat{\mathcal{L}}^d_{unif}$

B.7.2 Implementation in PyTorch

B.7.3 Datasets

CIFAR [180] We use the original training/test split with 50000 and 10000 images respectively of size 32×32.

STL-10 [63] In unsupervised pre-training, we use all labelled+unlabelled images (105000 images) for training and the remaining 8000 for test with size 96×96. During linear evaluation, we only use the 5000 training labelled images for learning the weights.

CUB200-2011 [294] This dataset is composed of 200 fine-grained bird species with 5994 training images and 5794 test images rescaled to 224×224.

Algorithm 2 Implementation in PyTorch

```
    # loader: generator of images
    # n: batch size
    # n_views: number of views
    # d: latent space dimension
    # f: encoder (with projection head)
    # x: Tensor of shape [n, *]
    # aug: augmentation module generating views
    # K: kernel defined on image space
    for x in loader:
      alphas = (K(x, x) + n*lamb*torch.eye(n)).inverse() @ K(x, x)
      x = aug(x, n_views) # shape=[n*n_views, *]
      z = f(x).view([n, n_views, d]) # shape=[n, n_views, d]
      mu = alphas.detach() @ z.mean(dim=1) # shape=[n, d]
      loss = L(mu)
      loss.backward()

    def L(mu, t=2):
      return torch.pdist(z, p=2).pow(2).mul(-t).exp().mean().log()

```

UTZappos [312] This dataset is composed of images of shoes from zappos.com. In order to be comparable with the literature on weakly supervised learning, we follow [280] and split it into 35017 training images and 15008 test images resized at 32×32.

ImageNet100 [74, 275] It is a subset of ImageNet containing 100 random classes and introduced in [275]. It contains 126689 training images and 5000 testing images rescaled to 224×224. It notably allows a reasonable computational time since we runt all our experiments on a single server node with 4 V100 GPU.

BHB [87] This dataset is composed of 10420 3D brain MRI images of size $121 \times 145 \times 121$ with $1.5mm^3$ spatial resolution. Only healthy subjects are included.

BIOBD [148] It is also a brain MRI dataset including 662 3D anatomical images and used for downstream classification. Each 3D volume has size $121 \times 145 \times 121$. It contains 306 patients with bipolar disorder vs 356 healthy controls and we aim at discriminating patients vs controls. It is particularly suited to investigate biomarkers discovery inside the brain [144].

CheXpert [154] This dataset is composed of 224 316 chest radiogaphs of 65240 patients. Each radiograph comes with 14 medical obervations. We use the official training set for our experiments, following [150, 154] and we test the models on the hold-out official validation split containing radiographs from 200 patients. For linear evaluation on this dataset, we train 5 linear probes to discriminate 5 pathologies (as binary classification) using only the radiographs with "certain" labels.

B.7.4 Contrastive models

Architecture. For all small-scale vision datasets (CIFAR-10 [180], CIFAR-100 [180], STL-10 [63], CUB200-2011 [294] and UT-Zappos [312]), we used official ResNet18 [135] backbone where we replaced the first 7×7 convolutional kernel by a smaller 3×3 kernel and we removed the first max-pooling layer for CIFAR-10, CIFAR-100 and UTZappos. For ImageNet100, we used ResNet50 [135] for stronger baselines as it is common in the literature. For medical images on brain MRI datasets (BHB [87] and BIOBD[148], we used DenseNet121 [149] as our default backbone encoder, following previous literature on these datasets [87].

Following [52], we use the representation space after the last average pooling layer with 2048 dimensions to perform linear evaluation and use a 2-layers MLP projection head with batch normalization between each layer for a final latent space with 128 dimensions.

Batch size. We always use a default batch size 256 for all experiments on vision datasets and 64 for brain MRI datasets (considering the computational cost with 3D images and since it had little impact on the performance [87]).

Optimization. We use SGD optimizer on small-scale vision datasets (CIFAR-10, CIFAR-100, STL-10, CUB200-2011, UT-Zappos) with a base learning rate $0.3 \times$ batch size$/256$ and a cosine scheduler. For ImageNet100, we use a LARS [311] optimizer with learning rate $0.02 \times \sqrt{\text{batch size}}$ and cosine scheduler. In Kernel Decoupled Uniformity loss, we set $\lambda = \frac{0.01}{\sqrt{\text{batch size}}}$ and $t = 2$. For SimCLR, we set the temperature to $\tau = 0.07$ for all datasets following [309]. Unless mentioned otherwise, we use 2 views for Decoupled Uniformity (both with and without kernel) and the computational cost remains comparable with standard contrastive models.

Training epochs. By default, we train the models for 200 epochs unless mentioned otherwise for all vision data-sets excepted CUB200-2011 and UTZappos where we train them for 1000 epochs, following [280]. For medical datasets, we perform pre-training for 50 epochs, as in [87]. For linear evaluation, we use a simple linear layer trained for 300 epochs with an initial learning rate 0.1 decayed by 0.1 on each plateau.

Augmentations. We follow [52] to define our full set of data augmentations for vision datasets including: *RandomResizedCrop* (uniform scale between 0.08 to 1), *RandomHorizontalFlip* and color distorsion (including color jittering and gray-scale). For medical datasets, we use cutout covering 25% of the image in each direction ($1/4^3$ of the entire volume), following [87].

Generative Models

Architecture. For VAE, we use ResNet18 backbone with a completely symmetric decoder using nearest-neighbor interpolation for up-sampling. For DCGAN, we follow the architecture described in [226]. We keep the original dimension for CIFAR-10 and CIFAR-100 datasets

and we resize the images to 64×64 for STL-10. For BigBiGAN [84], we use the ResNet50 pre-trained encoder available at https://tfhub.dev/deepmind/bigbigan-resnet50/1 with BN+CReLU features.

Training. For VAE, we use PyTorch-lightning pre-trained model for STL-10 [1] and we optimize VAE for CIFAR-10 and CIFAR-100 for 400 epochs using an initial learning rate 10^{-4} and SGD optimizer with a cosine scheduler. We use the same pipeline on RandBits dataset. For DCGAN, we optimize it using Adam optimizer (following [226]) and base learning rate 2×10^{-4}.

B.8 Omitted Proofs

B.8.1 Estimation Error with Empirical Decoupled Uniformity

Property 2. *$\hat{\mathcal{L}}^d_{unif}(f)$ fulfills $|\hat{\mathcal{L}}^d_{unif}(f) - \mathcal{L}^d_{unif}(f)| \leq O\left(\frac{1}{\sqrt{N}}\right)$.*

PROOF. For any $v \in \mathcal{V}$, since $f(x) \in \mathbb{S}^{d-1}$, then $||\mu_x|| = ||\mathbb{E}_{p_{\mathcal{A}}(v|x)} f(v)|| \leq \mathbb{E}_{p_{\mathcal{A}}(v|x)} ||f(v)|| = 1$. As a result, $e^{-||\mu_x - \mu_{x'}||^2} \in I \stackrel{\text{def}}{=} [e^{-4}, 1]$ for any $x, x' \in \mathcal{X}$. Since log is k-Lipschitz on I then:

$$|\hat{\mathcal{L}}^d_{unif}(f) - \mathcal{L}^d_{unif}(f)| \leq k \left| \frac{1}{N(N-1)} \sum_{i \neq j} e^{-||\mu_{x_i} - \mu_{x_j}||^2} - \mathbb{E}_{p(x)p(x')} e^{-||\mu_x - \mu_{x'}||^2} \right|$$

For a fixed $x \in \mathcal{X}$, let $g_N(x) = \frac{1}{N} \sum_{i=1}^N e^{-||\mu_x - \mu_{x_i}||^2}$ and $g(x) = \mathbb{E}_{p(x')} e^{-||\mu_x - \mu_{x'}||^2}$. Since $(Z_i)_{i \in [1..N]} = \left(e^{-||\mu_x - \mu_{x_i}||^2} - g(x) \right)_{i \in [1..N]}$ are iid with bounded support in $[-2, 2]$ and zero mean then by Berry–Esseen theorem we have $|g_N(x) - g(x)| \leq O(\frac{1}{\sqrt{N}})$. Similarly, $(Z'_i)_{i \in [1..N]} = \left(g_N(x_i) - \mathbb{E}_{p(x)} g_N(x) \right)$ are iid, bounded in $[-2, 2]$ and with zero mean. So $|\frac{1}{N} \sum_{i=1}^N g_N(x_i) - \mathbb{E}_{p(x)} g_N(x)| \leq O(\frac{1}{\sqrt{N}})$ by Berry–Esseen theorem. Then we have:

$$\begin{aligned}
|\hat{\mathcal{L}}^d_{unif}(f) - \mathcal{L}^d_{unif}(f)| &\leq k | \frac{N}{(N-1)N} \sum_{i=1}^N g_N(x_i) - \mathbb{E}_{p(x)} g(x)| \\
&\leq 2k | \frac{1}{N} \sum_{i=1}^N g_N(x_i) - \mathbb{E}_{p(x)} g_N(x) + \mathbb{E}_{p(x)} g_N(x) - \mathbb{E}_{p(x)} g(x)| \\
&\leq O(\frac{1}{\sqrt{N}}) + O(\frac{1}{\sqrt{N}}) \leq O(\frac{1}{\sqrt{N}})
\end{aligned}$$

B.8.2 Optimality of Decoupled Uniformity

Theorem 1. *(Optimality of Decoupled Uniformity) Given N points $(x_i)_{i \in [1..N]}$ such that $N \leq d + 1$, the optimal decoupled uniformity loss is reached when:*

1. *(Perfect uniformity) All centroids $(\mu_i)_{i \in [1..N]} = (\mu_{x_i})_{i \in [1..N]}$ make a regular simplex on the hyper-sphere $\mathbb{S}^{d-1}$*

[1] https://github.com/PyTorchLightning/pytorch-lightning

2. (Perfect alignment) f is perfectly aligned, i.e $\forall v, v' \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V|x_i), f(v) = f(v')$

PROOF. We will use Jensen's inequality and basic algebra to show these 2 properties. By triangular inequality, we have $||\mu_i|| = ||\mathbb{E}_{p_{\mathcal{A}}(v|x_i)} f(v)|| \leq \mathbb{E}||f(v)|| = 1$ since we assume $f(v) \in \mathbb{S}^d$. So all (μ_i) are bounded by 1.

Let $\mu = (\mu_i)_{i \in [1..N]}$. We have:

$$\begin{aligned}\Gamma(\mu) &:= \sum_{i,j=1}^{N} ||\mu_i - \mu_j||^2 = \sum_{i,j} ||\mu_i||^2 + ||\mu_j||^2 - 2\mu_i \cdot \mu_j \\ &\leq \sum_{i,j} (2 - 2\mu_i \cdot \mu_j) \\ &= 2N^2 - 2||\sum_i \mu_i||^2 \leq 2N^2\end{aligned}$$

with equality if and only if $\sum_{i=1}^{N} \mu_i = 0$ and $\forall i \in [1..N], ||\mu_i|| = 1$. By strict convexity of $u \to e^{-u}$, we have:

$$\begin{aligned}\sum_{i \neq j} \exp(-||\mu_i - \mu_j||^2) &\geq n(n-1) \exp\left(-\frac{\Gamma(\mu)}{n(n-1)}\right) \\ &\geq n(n-1) \exp\left(-\frac{2n}{n-1}\right)\end{aligned}$$

with equality if and only if all pairwise distance $||\mu_i - \mu_j||$ are equal (equality case in Jensen's inequality for strict convex function), $\sum_{i=1}^{N} \mu_i = 0$ and $||\mu_i|| = 1$. So all centroids must form a regular $n-1$-simplex inscribed on the hypersphere $\mathbb{S}^{d-1}$ centered at 0.

Finally, since $||\mu_i|| = 1$ then we have equality in the Jensen's inequality $||\mu_i|| = ||\mathbb{E}_{p_{\mathcal{A}}(v|x_i)} f(v)|| \leq \mathbb{E}_{p_{\mathcal{A}}(v|x_i)} ||f(v)|| = 1$. Since $||\cdot||$ is strictly convex on the hyper-sphere, then f must be constant on $\operatorname{supp} p_{\mathcal{A}}(\cdot|x_i)$, for all x_i so f must be perfectly aligned.

Theorem 2. *(Asymptotical Optimality) When the number of samples is infinite $N \to \infty$, then for any perfectly aligned encoder $f \in \mathcal{F}$ that minimizes $\mathcal{L}^d_{unif}$, the centroids μ_x for $x \sim p(X)$ are uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.*

PROOF. Let $f \in \mathcal{F}$ perfectly aligned. Then all centroids $\mu_x = f(x)$ lie on the hypersphere $\mathbb{S}^{d-1}$ and we are optimizing:

$$\arg\min_f \mathcal{L}^d_{unif}(f) = \arg\min_f \mathbb{E}_{x,x' \overset{iid}{\sim} p(X)} e^{-||f(x) - f(x')||^2}$$

So a direct application of Proposition 1. in [297] shows that the uniform distribution on $\mathbb{S}^{d-1}$ is the unique solution to this problem and that all centroids are uniformly distributed on the hyper-sphere.

B.8.3 Optimality of Supervised Loss

Lemma 6. *Let a downstream task $\mathcal{D}$ with C classes. We assume that $C \leq d+1$ (i.e., a big enough representation space), that all classes are balanced and the realizability of an encoder $f^* = \arg\min_{f \in \mathcal{F}} \mathcal{L}_{sup}(f)$ with $\mathcal{L}_{sup}(f) = \log \mathbb{E}_{y,y' \sim p(Y)p(Y')} e^{-||\mu_y - \mu_{y'}||^2}$, and $\mu_y = \mathbb{E}_{p(x|y)} \mu_x$. Then the optimal centroids $(\mu_y^*)_{y \in \mathcal{Y}}$ associated to f^* make a regular simplex on the hypersphere $\mathbb{S}^{d-1}$ and they are perfectly linearly separable, i.e $\min_{(w_y)_{y \in \mathcal{Y}} \in \mathbb{R}^d} \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{1}(w_y \cdot \mu_y^* < 0) = 0$.*

PROOF. This proof is very similar to the one in Theorem 2. We first notice that all "labelled" centroids $\mu_y = \mathbb{E}_{p(x|y)} \mu_x$ are bounded by 1 ($||\mu_y|| \leq \mathbb{E}_{p(x|y)} \mathbb{E}_{p_{\mathcal{A}}(v|x)} ||f(v)|| = 1$ by Jensen's inequality applied twice). Then, since all classes are balanced, we can re-write the supervised loss as:

$$\mathcal{L}_{sup}(f) = \log \frac{1}{C^2} \sum_{y,y'=1}^{C} e^{-||\mu_y - \mu_{y'}||^2}$$

We have:

$$\begin{aligned}
\Gamma_{\mathcal{Y}}(\mu) &:= \sum_{y,y'=1}^{C} ||\mu_y - \mu_{y'}||^2 = \sum_{y,y'} ||\mu_y||^2 + ||\mu_{y'}||^2 - 2\mu_y \cdot \mu_{y'} \\
&\leq \sum_{y,y'} (2 - 2\mu_y \cdot \mu_{y'}) \\
&= 2C^2 - 2||\sum_{y} \mu_y||^2 \leq 2C^2
\end{aligned}$$

with equality if and only if $\sum_{y=1}^{C} \mu_y = 0$ and $\forall y \in [1..C], ||\mu_y|| = 1$. By strict convexity of $u \to e^{-u}$, we have:

$$\begin{aligned}
\sum_{y \neq y'} \exp(-||\mu_y - \mu_{y'}||^2) &\geq C(C-1) \exp\left(-\frac{\Gamma_{\mathcal{Y}}(\mu)}{C(C-1)}\right) \\
&\geq C(C-1) \exp\left(-\frac{2C}{C-1}\right)
\end{aligned}$$

with equality if and only if all pairwise distance $||\mu_y - \mu_{y'}||$ are equal (equality case in Jensen's inequality for strict convex function), $\sum_{y=1}^{C} \mu_y = 0$ and $||\mu_y|| = 1$. So all centroids must form a regular $C-1$-simplex inscribed on the hypersphere $\mathbb{S}^{d-1}$ centered at 0. Furthermore, since $||\mu_y|| = 1$ then we have equality in the Jensen's inequality $||\mu_y|| = ||\mathbb{E}_{p(x|y)p_{\mathcal{A}}(v|x)} f(v)|| \leq \mathbb{E}_{p(x|y)p_{\mathcal{A}}(v|x)} ||f(v)|| = 1$ so f must by perfectly aligned for all samples belonging to the same class: $\forall x, x' \sim p(\cdot|y), f(x) = f(x')$.

B.8.4 Generalization bounds for decoupled uniformity

Theorem 7. *(Guarantees for a given downstream task) For any $f \in \mathcal{F}$ and augmentation distribution $\mathcal{A}$, we have:*

$$\mathcal{L}_{unif}^{d}(f) \leq \mathcal{L}_{unif}^{sup}(f) \leq 2\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y) + \mathcal{L}_{unif}^{d}(f) \leq 4\mathbb{E}_{p(x|y)p(x'|y)}||\mu_x - \mu_{x'}|| + \mathcal{L}_{unif}^{d}(f) \quad \text{(B.10)}$$

where $\mathrm{Var}(\mu_x^j|y) = \mathbb{E}_{p(x|y)}(\mu_x^j - \mathbb{E}_{p(x'|y)}\mu_{x'}^j)^2$ *and* μ_x^j *is the* j*-th component of* $\mu_x = \mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v)$.

PROOF.

Lower bound. To derive the lower bound, we apply Jensen's inequality to convex function $u \to e^{-u}$:

$$\begin{aligned}
\exp \mathcal{L}_{unif}^{d}(f) &= \mathbb{E}_{p(x)p(x')}e^{-||\mu_x - \mu_{x'}||^2} \\
&= \mathbb{E}_{p(x|y)p(x'|y)p(y)p(y')}e^{-||\mu_x - \mu_{x'}||^2} \\
&\leq \mathbb{E}_{p(y)p(y')} \exp\left(-\mathbb{E}_{p(x|y)p(x'|y')}||\mu_x - \mu_{x'}||^2\right)
\end{aligned}$$

Then, by Jensen's inequality applied to $||.||^2$:

$$\begin{aligned}
\mathbb{E}_{p(x|y)p(x'|y')}||\mu_x - \mu_{x'}||^2 &\overset{(1)}{=} \mathbb{E}_{p(x|y)}||\mu_x||^2 + \mathbb{E}_{p(x'|y')}||\mu_{x'}||^2 - 2\mu_y \cdot \mu_{y'} \\
&\geq ||\mathbb{E}_{p(x|y)}\mu_x||^2 + ||\mathbb{E}_{p(x'|y')}\mu_{x'}||^2 - 2\mu_y \cdot \mu_{y'} \\
&\overset{(1)}{=} ||\mu_y - \mu_{y'}||^2
\end{aligned}$$

(1) follows according to the previous lemma. So we can conclude:

$$\exp \mathcal{L}_{unif}^{d}(f) \leq \mathbb{E}_{p(y)p(y')} \exp(-||\mu_y - \mu_{y'}||^2) = \exp \mathcal{L}_{unif}^{sup}$$

Upper bound. For this bound, we will use the following equality (by definition of variance):

$$\begin{aligned}
||\mathbb{E}_{p(x|y)}\mu_x||^2 &= ||\mathbb{E}_{p(x|y)}\mu_x||^2 - \mathbb{E}_{p(x|y)}||\mu_x||^2 + \mathbb{E}_{p(x|y)}||\mu_x||^2 \\
&= -\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y) + \mathbb{E}_{p(x|y)}||\mu_x||^2
\end{aligned}$$

So we start by expending:

$$
\begin{aligned}
||\mu_y - \mu_{y'}||^2 &= ||\mathbb{E}_{p(x'|y')}\mu_{x'}||^2 + ||\mathbb{E}_{p(x|y)}\mu_x||^2 - 2\mathbb{E}_{p(x|y)p(x'|y')}\mu_x \cdot \mu_{x'} \\
&= \mathbb{E}_{p(x|y)}||\mu_x||^2 + \mathbb{E}_{p(x'|y')}||\mu_{x'}||^2 - \left(\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y) + \mathrm{Var}(\mu_{x'}^j|y)\right) - 2\mathbb{E}_{p(x|y)p(x'|y')}\mu_x \cdot \mu_{x'} \\
&= \mathbb{E}_{p(x|y)p(x'|y')}||\mu_x - \mu_{x'}||^2 - 2\left(\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y)\right)
\end{aligned}
$$

So by applying again Jensen's inequality:

$$
\begin{aligned}
\exp \mathcal{L}_{unif}^{sup} &= \mathbb{E}_{p(y)p(y')} \exp(-||\mu_y - \mu_{y'}||^2) \\
&\leq \mathbb{E}_{p(y)p(y')} \exp\left(-\mathbb{E}_{p(x|y)p(x'|y')}||\mu_x - \mu_{x'}||^2 + 2\left(\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y)\right)\right) \\
&\leq \exp 2\left(\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y_m)\right) \mathbb{E}_{p(y)p(y')} \exp\left(-\mathbb{E}_{p(x|y)p(x'|y')}||\mu_x - \mu_{x'}||^2\right) \\
&= \exp 2\left(\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y_m)\right) \exp \mathcal{L}_{unif}^{d}
\end{aligned}
$$

We set $y_m = \arg\max_{i,y \in [1..d] \times \mathcal{Y}} \mathrm{Var}(\mu_x^j|y)$ We conclude here by taking the log on the previous inequality.

Variance upper bound. Starting from the definition of conditional variance:

$$
\begin{aligned}
\sum_{j=1}^{d} \mathrm{Var}(\mu_x^j|y_m) &= \mathbb{E}_{p(x|y_m)}||\mu_x||^2 - ||\mathbb{E}_{p(x|y_m)}\mu_x||^2 \\
&= \mathbb{E}_{p(x|y_m)}\left((||\mu_x|| - ||\mathbb{E}_{p(x|y_m)}\mu_x||)(||\mu_x|| + ||\mathbb{E}_{p(x|y_m)}\mu_x||)\right) \\
&\overset{(1)}{\leq} \mathbb{E}_{p(x|y_m)}||\mu_x - \mathbb{E}_{p(x'|y_m)}\mu_{x'}||(||\mu_x|| + ||\mathbb{E}_{p(x|y_m)}\mu_x||) \\
&\overset{(2)}{\leq} 2\mathbb{E}_{p(x|y_m)}||\mu_x - \mathbb{E}_{p(x'|y_m)}\mu_{x'}|| \\
&\overset{(3)}{\leq} 2\mathbb{E}_{p(x|y_m)p(x'|y_m)}||\mu_x - \mu_{x'}||
\end{aligned}
$$

(1) Follows from standard inequality $||a - b|| \geq |||a|| - ||b|||$ (from Cauchy-Schwarz). (2) follows from boundness of $||\mu_x|| \leq 1$ and Jensen's inequality. (3) is again Jensen's inequality.

B.8.5 Generalization bound under intra-class connectivity assumption

Theorem 3. *Assuming 2, then for any ϵ-weak aligned encoder $f \in \mathcal{F}$:*

$$\mathcal{L}^d_{unif}(f) \leq \mathcal{L}^{sup}_{unif}(f) \leq 8D\epsilon + \mathcal{L}^d_{unif}(f) \tag{B.11}$$

Where D is the maximum diameter of all intra-class graphs G_y ($y \in \mathcal{Y}$).

PROOF. Let $y \in \mathcal{Y}$ and $x, x' \sim p(X|y)p(X'|y)$. By Assumption 2, it exists a path of length $p \leq D$ connecting (x, x') in G_y. So it exists $(x_i)_{i\in[1..p+1]} \in \mathcal{X}$ and $(v_i)_{i\in[1..p]} \in \mathcal{V}$ s.t $\forall i \in [1..p], v_i \sim p_{\mathcal{A}}(V|x_i) \cap p_{\mathcal{A}}(V|x_{i+1})$, $x_1 = x$ and $x_{p+1} = x'$. Then:

$$\begin{aligned}
||\mu_x - \mu_{x'}|| &= ||\mu_{x_1} - \mu_{x_{p+1}}|| \\
&= ||\sum_{i=1}^{p} \mu_{x_{i+1}} - \mu_{x_i}|| \\
&\leq \sum_{i=1}^{p} ||\mu_{x_{i+1}} - \mu_{x_i}|| \\
&= \sum_{i=1}^{p} ||\mu_{x_{i+1}} - f(v_i) + f(v_i) - \mu_{x_i}|| \\
&\leq \sum_{i=1}^{p} ||\mu_{x_{i+1}} - f(v_i)|| + ||f(v_i) - \mu_{x_i}|| \\
&\overset{(1)}{\leq} \sum_{i=1}^{p} \mathbb{E}_{p_{\mathcal{A}}(v|x_{i+1})}||f(v) - f(v_i)|| + \mathbb{E}_{p_{\mathcal{A}}(v|x_i)}||f(v_i) - f(v)|| \\
&\overset{(2)}{\leq} \sum_{i=1}^{p} (\epsilon + \epsilon) = 2\epsilon p \leq 2\epsilon D
\end{aligned}$$

(1) follows from Jensen's inequality and by definition of μ_x. (2) follows because f is ϵ-weak aligned and $v_i \sim p_{\mathcal{A}}(V|x_i) \cap p_{\mathcal{A}}(V|x_{i+1})$.

So we have $||\mu_x - \mu_{x'}|| \leq 2\epsilon D$ and we can conclude by Theorem 10 (right inequality).

B.8.6 Conditional Mean Embedding Estimation

Let $f \in \mathcal{F}$ fixed.

Theorem 4. *(Conditional Mean Embedding estimation) We assume that $\forall g \in \mathcal{H}_{\mathcal{X}}, \mathbb{E}_{p_{\mathcal{A}}(v|\cdot)}g(v) \in \mathcal{H}_{\mathcal{X}}$. Let $\{(v_1, x_1), ..., (v_N, x_N)\}$ iid samples from $p(V|X)p(X)$. Let $\Phi_N = [\phi(x_1), ..., \phi(x_N)]$ and $\Psi_f = [f(v_1), ..., f(v_N)]^T$. An estimator of the conditional mean embedding is:*

$$\forall x \in \mathcal{X}, \hat{\mu}_x = \sum_{i=1}^{N} \alpha_i(x) f(v_i) \tag{B.12}$$

where $\alpha_i(x) = \sum_{j=1}^{N}[(\Phi_N^T\Phi_N + \lambda N\mathbf{I}_N)^{-1}]_{ij}\langle\phi(x_j), \phi(x)\rangle_{\mathcal{H}_x}$. *It converges to* μ_x *with the* ℓ_2 *norm at a rate* $O(N^{-1/4})$ *for* $\lambda = O(\frac{1}{\sqrt{N}})$.

PROOF. Let $m_x = \mathbb{E}_{p_{\mathcal{A}}(v|x)}\langle f(v), f(\cdot)\rangle \in \mathcal{H}_{\mathcal{X}}$ be the conditional mean embedding operator. According to Theorem 6 in [263] and the assumption $\forall g \in \mathcal{H}_{\mathcal{X}}, \mathbb{E}_{p_{\mathcal{A}}(v|\cdot)}g(v) \in \mathcal{H}_{\mathcal{X}}$, this operator can be approximated by:

$$\hat{m}_x = \sum_{i=1}^{N}\alpha_i(x)\langle f(v_i), f(\cdot)\rangle$$

with α_i defined previously in the theorem. This estimator converges with RKHS norm to m_x at rate $O(\frac{1}{\sqrt{N\lambda}} + \lambda)$. So we need to link $m_x, \hat{m}_x$ with $\mu_x, \hat{\mu}_x$. We have:

$$\begin{aligned}
\langle m_x, \hat{m}_x\rangle_{\mathcal{H}_{\mathcal{X}}} &= \left\langle \mathbb{E}_{p_{\mathcal{A}}(v|x)}\langle f(v), f(\cdot)\rangle_{\mathbb{R}^d}, \sum_{i=1}^{N}\alpha_i(x)\langle f(v_i), f(\cdot)\rangle_{\mathbb{R}^d} \right\rangle_{\mathcal{H}_{\mathcal{X}}} \\
&= \sum_{i=1}^{N}\alpha_i(x)\left\langle \langle \mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v), f(\cdot)\rangle_{\mathbb{R}^d}, \langle f(v_i), f(\cdot)\rangle_{\mathbb{R}^d}\right\rangle_{\mathcal{H}_{\mathcal{X}}} \\
&\overset{(1)}{=} \sum_{i=1}^{N}\alpha_i(x)\langle \mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v), f(v_i)\rangle_{\mathbb{R}^d} \\
&= \langle \mu_x, \hat{\mu}_x\rangle_{\mathbb{R}^d}
\end{aligned}$$

(1) holds by the reproducing property of kernel $K_{\mathcal{X}}$ in $\mathcal{H}_{\mathcal{X}}$. We can similarly obtain:

$$\begin{aligned}
||m_x||^2_{\mathcal{H}_{\mathcal{X}}} &= \left\langle \mathbb{E}_{p_{\mathcal{A}}(v|x)}\langle f(v), f(\cdot)\rangle_{\mathbb{R}^d}, \mathbb{E}_{p_{\mathcal{A}}(v|x)}\langle f(v), f(\cdot)\rangle_{\mathbb{R}^d}\right\rangle_{\mathcal{H}_{\mathcal{X}}} \\
&\overset{(1)}{=} \langle \mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v), \mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v)\rangle_{\mathbb{R}^d} \\
&= ||\mathbb{E}_{p_{\mathcal{A}}(v|x)}f(v)||^2 = ||\mu_x||^2
\end{aligned}$$

Again, (1) by reproducing property of $K_{\mathcal{X}}$. And finally:

$$\begin{aligned}
||\hat{m}_x||^2_{\mathcal{H}_{\mathcal{X}}} &= \left\langle \sum_{i=1}^{N}\alpha_i(x)\langle f(v_i), f(\cdot)\rangle_{\mathbb{R}^d}, \sum_{i=1}^{n}\alpha_i(x)\langle f(v_i), f(\cdot)\rangle_{\mathbb{R}^d}\right\rangle_{\mathcal{H}_{\mathcal{X}}} \\
&= \sum_{i,j}\alpha_i(x)\alpha_j(x)\langle f(v_i), f(v_j)\rangle_{\mathbb{R}^d} \\
&= ||\hat{\mu}_x||^2_{\mathbb{R}^d}
\end{aligned}$$

By pooling these 3 equalities, we have:

$$\begin{aligned}
||m_x - \hat{m}_x||^2_{\mathcal{H}_{\mathcal{X}}} &= ||m_x||^2 + ||\hat{m}_x||^2 - 2\langle m_x, \hat{m}_x\rangle \\
&= ||\mu_x||^2 + ||\hat{\mu}_x||^2 - 2\langle \mu_x, \hat{\mu}_x\rangle \\
&= ||\mu_x - \hat{\mu}_x||^2_{\mathbb{R}^d}
\end{aligned}$$

We can conclude since $||m_x - \hat{m}_x|| \leq O(\lambda + (N\lambda)^{-1/2})$.

B.8.7 Generalization bound under extended intra-class connectivity hypothesis

Theorem. *Assuming 4 and 3 holds for a reproducible kernel $K_{\mathcal{X}}$ and augmentation distribution $\mathcal{A}$. Let $f \in \mathcal{F}$ α-aligned. Let $(x_i)_{i\in[1..N]} \overset{i.i.d.}{\sim} p(X)$. We have:*

$$\mathcal{L}^d_{unif}(f) \leq \mathcal{L}^{sup}_{unif}(f) \leq \mathcal{L}^d_{unif}(f) + 4D(2\alpha + \beta_N(K_{\mathcal{X}})\epsilon) + O(N^{-1/4}) \tag{B.13}$$

where $\beta_N(K_{\mathcal{X}}) = (\frac{\lambda_{min}(K_N)}{\sqrt{N}} + \sqrt{N}\lambda)^{-1} = O(1)$ for $\lambda = O(\frac{1}{\sqrt{N}})$, $K_N = (K_{\mathcal{X}}(x_i, x_j))_{i,j\in[1..N]}$ and D is the maximal diameter for all $\tilde{G}_y$, $y \in \mathcal{Y}$. We noted $\lambda_{min}(K_N)$ is the minimal eigenvalue of K_N.

PROOF. Let $y \in \mathcal{Y}$ and $x, x' \sim p(X|Y = y)p(X'|Y = y)$. By Assumption 3, it exists a path of length $p \leq D$ connecting x, x' in $\tilde{G}$. So it exists $(\bar{u}_i)_{i\in[1..p+1]} \in \mathcal{X}$ and $(u_i)_{i\in I} \in \mathcal{V}$ s.t $\forall i \in I, u_i \sim p_{\mathcal{A}}(V|\bar{u}_i) \cap p_{\mathcal{A}}(V|\bar{u}_{i+1})$ and $\forall j \in J, \max(K(\bar{u}_j, \bar{u}_j), K(\bar{u}_{j+1}, \bar{u}_{j+1})) - K(\bar{u}_j, \bar{u}_{j+1}) \leq \epsilon$ with (I, J) a partition of $[1..p]$. Furthermore, $\bar{u}_1 = x$ and $\bar{u}_{p+1} = x'$. As a result, we have:

$$\begin{aligned}
||\mu_x - \mu_{x'}|| &= ||\mu_{\bar{u}_1} - \mu_{\bar{u}_p}|| \\
&= ||\sum_{i=1}^{p} \mu_{\bar{u}_{i+1}} - \mu_{\bar{u}_i}|| \\
&\leq \sum_{i=1}^{p} ||\mu_{\bar{u}_{i+1}} - \mu_{\bar{u}_i}|| \\
&= \sum_{i\in I} ||\mu_{\bar{u}_{i+1}} - \mu_{\bar{u}_i}|| + \sum_{j\in J} ||\mu_{\bar{u}_{j+1}} - \mu_{\bar{u}_j}||
\end{aligned}$$

Edges in E. As in proof of Theorem 4, we use the α-alignment of f to derive a bound:

$$\begin{aligned}
\sum_{i\in I} ||\mu_{\bar{u}_{i+1}} - \mu_{\bar{u}_i}|| &= \sum_{i\in I} ||\mu_{\bar{u}_{i+1}} - f(u_i) + f(u_i) - \mu_{\bar{u}_i}|| \\
&\leq \sum_{i\in I} ||\mu_{\bar{u}_{i+1}} - f(u_i)|| + ||f(u_i) - \mu_{\bar{u}_i}|| \\
&\overset{(1)}{\leq} \sum_{i\in I} \mathbb{E}_{p_{\mathcal{A}}(u|\bar{u}_{i+1})} ||f(u) - f(u_i)|| + \mathbb{E}_{p_{\mathcal{A}}(u|\bar{u}_i)} ||f(u_i) - f(u)|| \\
&\overset{(2)}{\leq} \sum_{i\in I} (\alpha + \alpha) = 2\alpha|I|
\end{aligned}$$

(1) holds by Jensen's inequality and (2) because f is α-aligned.

Edges in E_K For this bound, we will use Theorem 5 to approximate $\mu_{\bar{u}}$ and then derive a bound from the property of G^{ϵ}_K. Let $v_k \sim p_{\mathcal{A}}(V|x_k)$ for $k \in [1..N]$. By Theorem 5, we

know that, for all $j \in J$, $\hat{\mu}_{\bar{u}_j}$ converges to $\mu_{\bar{u}_j}$ with ℓ_2 norm at rate $O(N^{-1/4})$ where $\hat{\mu}_{\bar{u}_j} = \sum_{k,l=1}^{N} \alpha_{k,l} K_{\mathcal{X}}(x_l, \bar{u}_j) f(v_k)$ and $\alpha_{k,l} = [(K_N + N\lambda \mathbf{I}_N)^{-1}]_{k,l}$. As a result, for any $j \in J$, we have:

$$\begin{aligned}
||\mu_{\bar{u}_{j+1}} - \mu_{\bar{u}_j}|| &= ||\mu_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_{j+1}} + \hat{\mu}_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_j} + \hat{\mu}_{\bar{u}_j} - \mu_{\bar{u}_j}|| \\
&\leq ||\mu_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_{j+1}}|| + ||\hat{\mu}_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_j}|| + ||\hat{\mu}_{\bar{u}_j} - \mu_{\bar{u}_j}|| \\
&\overset{(1)}{\leq} O\left(\frac{1}{N^{1/4}}\right) + ||\hat{\mu}_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_j}||
\end{aligned}$$

Where (1) holds by Theorem 5. Then we will need the following lemma to conclude:

Lemma. For any $a, b, c \in \mathcal{X}, \max(K(a,a), K(b,b)) - K(a,b) \geq |K(a,c) - K(b,c)|$ for any reproducible kernel K.

PROOF. Let $a, b, c \in \mathcal{X}$. We consider the distance $d(x,y) = K(x,x) + K(y,y) - 2K(x,y)$ (it is a distance since K is a reproducible kernel so it can be expressed as $K(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$). We will distinguish two cases.

Case 1. We assume $K(a,c) \geq K(b,c)$. We have the following triangular inequality:

$$\begin{aligned}
& d(a,b) + d(a,c) \geq d(b,c) \\
\implies & K(a,b) + K(b,b) - 2K(a,b) + K(a,a) + K(c,c) - 2K(a,c) \geq K(b,b) + K(c,c) - 2K(b,c) \\
\implies & K(a,a) - K(a,b) \geq K(a,c) - K(b,c) \geq 0
\end{aligned}$$

So $\max(K(a,a), K(b,b)) - K(a,b) \geq |K(a,c) - K(b,c)|$.

Case 2. We assume $K(b,c) \geq K(a,c)$. We apply symmetrically the triangular inequality:

$$\begin{aligned}
& d(a,b) + d(b,c) \geq d(a,c) \\
\implies & K(b,b) - K(a,b) \geq K(b,c) - K(a,c) \geq 0
\end{aligned}$$

So $\max(K(a,a), K(b,b)) - K(a,b) \geq |K(a,c) - K(b,c)|$, concluding the proof.

Then, by definition of $\hat{\mu}_{\bar{u}_j}$:

$$\begin{aligned}
||\hat{\mu}_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_j}|| &= ||\sum_{k,l=1}^{N} \alpha_{k,l} K(x_l, \bar{u}_{j+1}) f(v_k) - \sum_{k,l=1}^{N} \alpha_{k,l} K(x_l, \bar{u}_j) f(v_k)|| \\
&= ||AC||
\end{aligned}$$

Where $A = (\sum_{k=1}^{N} \alpha_{kj} f(v_k)^i)_{i,j} \in \mathbb{R}^{d \times N}$ ($f(\cdot)^i$ is the i-th component of $f(\cdot)$) and $C =$

$(K(x_l, \bar{u}_{j+1}) - K(x_l, \bar{u}_j))_l \in \mathbb{R}^{n\times 1}$. So, using the property of spectral ℓ_2 norm we have:

$$||\hat{\mu}_{\bar{u}_{j+1}} - \hat{\mu}_{\bar{u}_j}|| = ||AC|| \leq ||A||_2||C||_2$$

Using the previous lemma and because $(\bar{u}_j, \bar{u}_{j+1}) \in E_K$, we have: $||C||_2^2 = \sum_{i=1}^N (K(x_i, \bar{u}_{j+1}) - K(x_i, \bar{u}_j))^2 \leq \sum_{i=1}^N (\max(K(\bar{u}_{j+1}, \bar{u}_{j+1}), K(\bar{u}_j, \bar{u}_j)) - K(\bar{u}_j, \bar{u}_{j+1}))^2 \leq N\epsilon^2$. To conclude, we will prove that $||A||_2 \leq ||\alpha||_2$ where $\alpha = (\alpha_{ij})_{i,j\in[1..N]^2}$. For any $v \in \mathbb{R}^N$, we have:

$$||Av||^2 = ||\sum_{k,j=1}^n \alpha_{k,j} v_j f(x_k)||^2 \overset{(1)}{\leq} \left(\sum_{k,j=1}^n \alpha_{k,j} v_j\right)^2 = ||\alpha v||^2 \overset{(2)}{\leq} ||\alpha||_2^2 ||v||^2$$

Where (1) holds with Cauchy-Schwarz inequality and because $f(\cdot) \in \mathbb{S}^{d-1}$ and (2) holds by definition of spectral ℓ_2 norm. So we have $\forall v \in \mathbb{R}^d, ||Av|| \leq ||\alpha||_2||v||$, showing that $||A||_2 \leq ||\alpha||_2$.

So we can conclude that:

$$\begin{aligned}\sum_{j\in J} ||\mu_{\bar{u}_{j+1}} - \mu_{\bar{u}_j}|| &\leq \sum_{j\in J} \left(\sqrt{N}||(K_N + \lambda N \mathbf{I}_N)^{-1}||_2\epsilon + O(N^{-1/4})\right) \\ &= |J|||(K_N + N\lambda\mathbf{I}_N)^{-1}||_2\sqrt{N}\epsilon + O(N^{-1/4})\end{aligned}$$

We set $\beta_N(K_N) = \sqrt{N}||(K_N + \lambda N\mathbf{I}_N)^{-1}||_2$. In order to see that $\beta_N(K_N) = (\frac{\lambda_{min}(K_N)}{\sqrt{N}} + \sqrt{N}\lambda)^{-1}$ with $\lambda_{min}(K_N) > 0$ the minimum eigenvalue of K_N, we apply the spectral theorem on the symmetric definite-positive kernel matrix K_N. Let $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$ the eigenvalues of K_N. According to the spectral theorem, it exists U an unitary matrix such that $K_N = UDU^T$ with $D = \text{diag}(\lambda_1, ..., \lambda_N)$. So, by definition of spectral norm:

$$\begin{aligned}||(K_N + N\lambda\mathbf{I}_N)^{-1}||_2^2 &= \lambda_{max}\left(U(D + N\lambda\mathbf{I}_N)^{-1}U^T U(D + \lambda N\mathbf{I}_N)^{-1}U^T\right) \\ &= \lambda_{max}(U\tilde{D}U^T) \\ &= (\lambda_1 + N\lambda)^{-2}\end{aligned}$$

where $\tilde{D} = \text{diag}(\frac{1}{(\lambda_1+N\lambda)^2}, ..., \frac{1}{(\lambda_N+N\lambda)^2})$. So we can conclude that $\beta_N(K_N) = (\frac{\lambda_1}{\sqrt{N}} + \sqrt{N}\lambda)^{-1} = O(1)$ for $\lambda = O(\frac{1}{\sqrt{N}})$.

Finally, by pooling inequalities for edges over E and E_K, we have:

$$||\mu_x - \mu_{x'}|| \leq 2\alpha|I| + |J|\beta_N(K_N)\epsilon + O(N^{-1/4}) \leq D(2\alpha + \beta_N(K_N)\epsilon) + O(N^{-1/4})$$

We can conclude by plugging this inequality in Theorem 10.

Theorem 5. *We assume 3 and 4 hold for a reproducible kernel $K_\mathcal{X}$ and augmentation module*

$\mathcal{A}$. *Let* $(v_i, x_i)_{i\in[1..N]} \overset{i.i.d.}{\sim} p_{\mathcal{A}}(V, X)$. *Let* $\hat{\mu}_{x_j} = \sum_{i=1}^{N} \alpha_{i,j} f(v_i)$ *with* $\alpha_{i,j} = ((K_N + \lambda \mathbf{I}_N)^{-1} K_N)_{ij}$ *and* $K_N = [K_{\mathcal{X}}(x_i, x_j)]_{i,j\in[1..N]}$. *Then the empirical decoupled uniformity loss,*

$$\hat{\mathcal{L}}^d_{unif} \overset{\text{def}}{=} \log \frac{1}{N(N-1)} \sum_{i,j=1}^{N} \exp(-||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2)$$

verifies, for any α*-weak aligned encoder* $f \in \mathcal{F}$*:*

$$\hat{\mathcal{L}}^d_{unif} - O\left(\frac{1}{N^{1/4}}\right) \leq \mathcal{L}^{sup}_{unif}(f) \leq \hat{\mathcal{L}}^d_{unif} + 4D(2\alpha + \beta_N(K_{\mathcal{X}})\epsilon) + O\left(\frac{1}{N^{1/4}}\right) \tag{B.14}$$

PROOF. We just need to prove that, for any $f \in \mathcal{F}$, $|\mathcal{L}^d_{unif}(f) - \hat{\mathcal{L}}^d_{unif}(f)| \leq O(N^{-1/4})$ and we can conclude through the previous theorem. We have:

$$\begin{aligned}
|\mathcal{L}^d_{unif}(f) - \hat{\mathcal{L}}^d_{unif}(f)| &= \left| \log \frac{1}{N(N-1)} \sum_{i,j=1}^{N} \exp(-||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2) - \mathbb{E}_{p(x)p(x')} e^{-||\mu_x - \mu_{x'}||^2} \right| \\
&\leq \underbrace{\left| \log \frac{1}{N(N-1)} \sum_{i,j=1}^{N} \exp(-||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2) - \log \frac{1}{N(N-1)} e^{-||\mu_{x_i} - \mu_{x_j}||^2} \right|}_{=E_1} \\
&+ \left| \log \frac{1}{N(N-1)} e^{-||\mu_{x_i} - \mu_{x_j}||^2} - \mathbb{E}_{p(x)p(x')} e^{-||\mu_x - \mu_{x'}||^2} \right|
\end{aligned}$$

The second term in last inequality is bounded by $O(\frac{1}{\sqrt{N}})$ according to property 2. As for the first term, we use the fact that log is k-Lipschitz continuous on $[e^{-4}, 1]$ and exp is k'-Lipschitz continuous on $[-4, 0]$ so:

$$\begin{aligned}
|E_1| &\leq \frac{k}{N(N-1)} \left| \sum_{i,j=1}^{N} e^{-||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2} - e^{-||\mu_{x_i} - \mu_{x_j}||^2} \right| \\
&\leq \frac{kk'}{N(N-1)} \left| \sum_{i,j=1}^{N} ||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2 - ||\mu_{x_i} - \mu_{x_j}||^2 \right|
\end{aligned}$$

Finally, we conclude using the boundness of $\hat{\mu}_x$ and μ_x by a constant C:

$$\begin{aligned}
||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}||^2 - ||\mu_{x_i} - \mu_{x_j}||^2 &= (||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}|| + ||\mu_{x_i} - \mu_{x_j}||)(||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}|| - ||\mu_{x_i} - \mu_{x_j}||) \\
&\leq 4C(||\hat{\mu}_{x_i} - \hat{\mu}_{x_j}|| - ||\mu_{x_i} - \mu_{x_j}||) \\
&\leq 4C||\hat{\mu}_{x_i} - \hat{\mu}_{x_j} - (\mu_{x_i} - \mu_{x_j})|| \\
&\leq 4C(||\hat{\mu}_{x_i} - \mu_{x_i}|| + ||\hat{\mu}_{x_j} - \mu_{x_j}||) \\
&= O\left(\frac{1}{N^{-1/4}}\right)
\end{aligned}$$

Appendix C

Third Appendix

C.1 Theoretical comparison with EnD

Here, we present a more detailed theoretical analysis of EnD [273] and we show the regularization term is equivalent to conditions (4.7) and another symmetric condition on negative samples:

$$a)\frac{1}{P_c}\sum_k s_k^{+,b'} - \frac{1}{P_a}\sum_i s_i^{+,b} = 0 \quad b)\frac{1}{N_c}\sum_t s_t^{-,b'} - \frac{1}{N_a}\sum_j s_j^{-,b} = 0 \tag{C.1}$$

which can be turned into a minimization term $\mathcal{R}$, using the method of Lagrange multipliers:

$$\mathcal{R} = -\lambda_1\left(\frac{1}{P_c}\sum_k s_k^{+,b'} - \frac{1}{P_a}\sum_i s_i^{+,b}\right) - \lambda_2\left(\frac{1}{N_c}\sum_t s_t^{-,b'} - \frac{1}{N_a}\sum_j s_j^{-,b}\right) \tag{C.2}$$

Now, if we assume $\lambda_1 = \lambda_2 = 1$, we can re-arrange the terms, obtaining:

$$\mathcal{R} = \underbrace{\left(\frac{1}{P_a}\sum_i s_i^{+,b} + \frac{1}{N_a}\sum_j s_j^{-,b}\right)}_{\mathcal{R}^{\perp}} - \underbrace{\left(\frac{1}{P_c}\sum_k s_k^{+,b'} + \frac{1}{N_c}\sum_t s_t^{-,b'}\right)}_{\mathcal{R}^{\parallel}} \tag{C.3}$$

The two terms we obtain are equivalent to the disentangling term $\mathcal{R}^{\perp}$ and to the entangling term $\mathcal{R}^{\parallel}$ of the EnD techniques [273]: $\mathcal{R}^{\perp}$ tries to decorrelate all of the samples which share the same bias attribute, while the $\mathcal{R}^{\parallel}$ tries to maximize the correlation of samples which belong to the same class but have different bias attributes. Notably, the $\mathcal{R}^{\perp}$ also employs the absolute values, in order to avoid anti-correlating bias-aligned samples.

Bibliography

[1] A. Abrol, H. Rokham, and V. D. Calhoun. Diagnostic and prognostic classification of brain disorders using residual learning on structural mri data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4084–4088. IEEE, 2019.

[2] A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17, 2021.

[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

[4] A. Aglinskas, J. K. Hartshorne, and S. Anzellotti. Contrastive machine learning reveals the structure of neuroanatomical variation within autism. *Science*, 376(6597):1070–1074, 2022.

[5] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[6] Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[7] M. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[8] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1): 1–74, 2021.

[9] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165, 2017.

[10] K. Armanious, S. Abdulatif, W. Shi, S. Salian, T. Küstner, D. Weiskopf, T. Hepp, S. Gatidis, and B. Yang. Age-net: An mri-based iterative framework for biological age estimation. *arXiv preprint arXiv:2009.10765*, 2020.

[11] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. *arXiv:1902.09229 [cs, stat]*, Feb. 2019. URL http://arxiv.org/abs/1902.09229. arXiv: 1902.09229.

[12] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1): 95–113, 2007.

[13] B. B. Avants, N. Tustison, and G. Song. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.

[14] M. Awad and R. Khanna. Support vector regression. In *Efficient learning machines*, pages 67–80. Springer, 2015.

[15] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

[16] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, M. Uhlig, A. Anwander, P.-L. Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6 (1):1–21, 2019.

[17] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020.

[18] J. T. Baker, A. J. Holmes, G. A. Masters, B. T. Yeo, F. Krienen, R. L. Buckner, and D. Öngür. Disruption of cortical association networks in schizophrenia and psychotic bipolar disorder. *JAMA psychiatry*, 71(2):109–118, 2014.

[19] R. Balestriero, L. Bottou, and Y. LeCun. The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*, 2022.

[20] G. Ball, C. E. Kelly, R. Beare, and M. L. Seal. Individual variation underlying brain age estimates in typical development. *NeuroImage*, 235:118036, 2021.

[21] C. A. Barbano, E. Tartaglione, and M. Grangetto. Bridging the gap between debiasing and privacy for deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3806–3815, 2021.

[22] A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[23] H. B. Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

[24] V. M. Bashyam, J. Doshi, G. Erus, D. Srinivasan, A. Abdulkadir, M. Habes, Y. Fan, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, T. D. Satterthwaite, D. H. Wolf, R. E. Gur, R. C. Gur, J. C. Morris, M. S. Albert, H. J. Grabe, S. M. Resnick, R. N. Bryan, D. A. Wolk, H. Shou, I. M. Nasrallah, and C. Davatzikos. Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging, 2020.

[25] V. M. Bashyam, G. Erus, J. Doshi, M. Habes, I. Nasralah, M. Truelove-Hill, D. Srinivasan, L. Mamourian, R. Pomponio, Y. Fan, et al. Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143 (7):2312–2324, 2020.

[26] V. M. Bashyam, G. Erus, J. Doshi, M. Habes, I. M. Nasrallah, M. Truelove-Hill, D. Srinivasan, L. Mamourian, R. Pomponio, Y. Fan, et al. Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143(7):2312–2324, 2020.

[27] S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[28] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.

[29] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[30] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[31] C. M. Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[32] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[33] S. V. Borodachov, D. P. Hardin, and E. B. Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

[34] V. Borrell. How cells fold the cerebral cortex. *Journal of Neuroscience*, 38(4):776–783, 2018.

[35] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific reports*, 10(1):1–16, 2020.

[36] R. L. Buckner, J. L. Roffman, and J. W. Smoller. Brain genomics superstruct project (gsp), 2014. URL https://doi.org/10.7910/DVN/25833.

[37] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[38] D. Bzdok and A. Meyer-Lindenberg. Machine learning for precision psychiatry. *arXiv preprint arXiv:1705.10553*, 2017.

[39] D. Bzdok and A. Meyer-Lindenberg. Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230, 2018.

[40] D. Bzdok, D. L. Floris, and A. F. Marquand. Analysing brain networks in population neuroscience: a case for the bayesian philosophy. *Philosophical Transactions of the Royal Society B*, 375(1796):20190661, 2020.

[41] R. Cadene, C. Dancette, M. Cord, D. Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32: 841–852, 2019.

[42] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[43] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[44] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[45] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, pages 1–36, 2021.

[46] C. Chadebec, E. Thibeau-Sutre, N. Burgos, and S. Allassonnière. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[47] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33, 2020.

[48] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, pages 416–422, 2001.

[49] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019.

[50] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101539. URL http://www.sciencedirect.com/science/article/pii/S1361841518304699.

[51] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[53] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[54] T. Chen, C. Luo, and L. Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

[55] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[56] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[57] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[58] C. R. Ching, D. P. Hibar, T. P. Gurholt, A. Nunes, S. I. Thomopoulos, C. Abé, I. Agartz, R. M. Brouwer, D. M. Cannon, S. M. de Zwarte, et al. What we learn about bipolar

disorder from large-scale neuroimaging: Findings and future directions from the enigma bipolar disorder working group. *Human brain mapping*, 43(1):56–82, 2022.

[59] N. Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

[60] S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005. ISBN 978-0-7695-2372-9. doi: 10.1109/CVPR.2005.202. URL http://ieeexplore.ieee.org/document/1467314/.

[61] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran Associates, Inc., 2020.

[62] C. Clark, M. Yatskar, and L. Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[63] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[64] J. H. Cole and K. Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017.

[65] J. H. Cole and K. Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017.

[66] J. H. Cole, R. Leech, D. J. Sharp, and A. D. N. Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, 77(4): 571–581, 2015.

[67] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[68] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. V. Hernández, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

[69] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. V. Hernández, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

[70] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[71] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[72] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[73] S. Dehaene, F. Al Roumi, Y. Lakretz, S. Planton, and M. Sablé-Meyer. Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 2022.

[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[75] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[76] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[77] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, et al. Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170, 2019.

[78] B. E. Dewey, L. Zuo, A. Carass, Y. He, Y. Liu, E. M. Mowry, S. Newsome, J. Oh, P. A. Calabresi, and J. L. Prince. A disentangled latent space for cross-site mri harmonization. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 720–729, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59728-3. doi: 10.1007/978-3-030-59728-3_70.

[79] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[80] A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1):1–15, 2017.

[81] M. Dinomais, S. Celle, G. T. Duval, F. Roche, S. Henni, R. Bartha, O. Beauchet, and C. Annweiler. Anatomic correlation of the mini-mental state examination: a voxel-based morphometric study in older adults. *PloS one*, 11(10):e0162889, 2016.

[82] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage*, 228:117689, 2021.

[83] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[84] J. Donahue and K. Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019.

[85] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[86] B. Dufumier, P. Gori, I. Battaglia, J. Victor, A. Grigis, and E. Duchesnay. Benchmarking cnn on 3d anatomical brain mri: Architectures, data augmentation and deep ensemble learning. *arXiv preprint arXiv:2106.01132*, 2021.

[87] B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, P. Favre, M. Polosan, C. McDonald, C. M. Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International conference on medical image computing and computer-assisted intervention*. Springer, 2021.

[88] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[89] L. Eliot, A. Ahmed, H. Khan, and J. Patel. Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neuroscience & Biobehavioral Reviews*, 125:667–697, 2021.

[90] D. Erhan, A. Courville, and Y. Bengio. Understanding representations learned in deep architectures. Technical report, Université de Montréal/DIRO, 2010.

[91] L. Ericsson, H. Gouk, and T. M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[92] T. Eslami, F. Almuqhim, J. S. Raiker, and F. Saeed. Machine learning methods for diagnosing autism spectrum disorder and attention-deficit/hyperactivity disorder using functional and structural mri: A survey. *Frontiers in neuroinformatics*, 14:62, 2021.

[93] F. Farokhian, I. Beheshti, D. Sone, and H. Matsuda. Comparing cat12 and vbm8 for detecting brain morphological abnormalities in temporal lobe epilepsy. *Frontiers in neurology*, 8:428, 2017.

[94] A. Fedorov, L. Wu, T. Sylvain, M. Luck, T. P. DeRamus, D. Bleklov, S. M. Plis, and V. D. Calhoun. On self-supervised multimodal representation learning: an application to alzheimer's disease. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1548–1552. IEEE, 2021.

[95] A. Filos, S. Farquhar, A. N. Gomez, T. G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

[96] L. Fisch, R. Leenings, N. R. Winter, U. Dannlowski, C. Gaser, J. H. Cole, and T. Hahn. Predicting chronological age from structural neuroimaging: The predictive analytics competition 2019. *Frontiers in Psychiatry*, 12, 2021.

[97] B. Fischl, M. I. Sereno, and A. M. Dale. Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[98] B. Fischl, A. Van Der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, et al. Automatically parcellating the human cerebral cortex. *Cerebral cortex*, 14(1):11–22, 2004.

[99] C. Flint, M. Cearns, N. Opel, R. Redlich, D. Mehler, D. Emden, N. R. Winter, R. Leenings, S. B. Eickhoff, T. Kircher, et al. Systematic overestimation of machine learning performance in neuroimaging studies of depression. *arXiv preprint arXiv:1912.06686*, 2019.

[100] D. J. Follmer, S.-Y. Fang, R. B. Clariana, B. J. Meyer, and P. Li. What predicts adult readers' understanding of stem texts? *Reading and Writing*, 31(1):185–214, 2018.

[101] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

[102] A. Foster, R. Pukdee, and T. Rainforth. Improving transformation invariance in contrastive representation learning. *arXiv preprint arXiv:2010.09515*, 2020.

[103] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[104] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[105] Y. Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

[106] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[107] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. In *Advances in neural information processing systems*, pages 3581–3590, 2017.

[108] R. Garcia-Dias, C. Scarpazza, L. Baecker, S. Vieira, W. H. Pinaya, A. Corvin, A. Redolfi, B. Nelson, B. Crespo-Facorro, C. McDonald, D. Tordesillas-Gutiérrez, D. Cannon, D. Mothersill, D. Hernaus, D. Morris, E. Setien-Suero, G. Donohoe, G. Frisoni, G. Tronchin, J. Sato, M. Marcelis, M. Kempton, N. E. van Haren, O. Gruber, P. McGorry, P. Amminger, P. McGuire, Q. Gong, R. S. Kahn, R. Ayesa-Arriola, T. van Amelsvoort, V. Ortiz-García de la Foz, V. Calhoun, W. Cahn, and A. Mechelli. Neuroharmony: A new tool for harmonizing volumetric mri data from unseen scanners. *NeuroImage*, 220: 117127, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117127. URL https://www.sciencedirect.com/science/article/pii/S1053811920306133.

[109] C. Gaser and R. Dahnke. Cat-a computational anatomy toolbox for the analysis of structural mri data. *HBM*, 2016:336–348, 2016.

[110] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, and A. D. N. Initiative. Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease. *PloS one*, 8(6):e67346, 2013.

[111] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[112] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

[113] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[114] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[115] D. Goldman. On dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *arXiv*, 2021.

[116] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[117] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[118] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[119] D. N. Greve, L. Van der Haegen, Q. Cai, S. Stufflebeam, M. R. Sabuncu, B. Fischl, and M. Brysbaert. A surface-based analysis of language lateralization and cortical asymmetry. *Journal of cognitive neuroscience*, 25(9):1477–1492, 2013.

[120] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[121] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[122] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

[123] F. K. Gustafsson, M. Danelljan, and T. B. Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[124] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[125] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

[126] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[127] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.

[128] S. V. Haijma, N. Van Haren, W. Cahn, P. C. M. Koolschijn, H. E. Hulshoff Pol, and R. S. Kahn. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia bulletin*, 39(5):1129–1138, 2013.

[129] T. Hajek, E. Gunde, C. Slaney, L. Propper, G. MacQueen, A. Duffy, and M. Alda. Amygdala and hippocampal volumes in relatives of patients with bipolar disorder: A high—risk study. *The Canadian Journal of Psychiatry*, 54(11):726–733, 2009.

[130] T. Hajek, M. Kopecek, C. Höschl, and M. Alda. Smaller hippocampal volumes in patients with bipolar disorder are masked by exposure to lithium: a meta-analysis. *Journal of Psychiatry and Neuroscience*, 37(5):333–343, 2012.

[131] L. K. Han, R. Dinga, T. Hahn, C. R. Ching, L. T. Eyler, L. Aftanas, M. Aghajani, A. Aleman, B. T. Baune, K. Berger, et al. Brain aging in major depressive disorder: results from the enigma major depressive disorder working group. *Molecular psychiatry*, pages 1–16, 2020.

[132] L. C. Hanford, A. Nazarov, G. B. Hall, and R. B. Sassi. Cortical thickness in bipolar disorder: a systematic review. *Bipolar disorders*, 18(1):4–18, 2016.

[133] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*, 2021.

[134] J. Z. HaoChen, C. Wei, A. Kumar, and T. Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

[135] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[136] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[137] S. He, P. E. Grant, and Y. Ou. Global-local transformer for brain age estimation. *IEEE Transactions on Medical Imaging*, 41(1):213–224, 2021.

[138] S. He, D. Pereira, J. D. Perez, R. L. Gollub, S. N. Murphy, S. Prabhu, R. Pienaar, R. L. Robertson, P. E. Grant, and Y. Ou. Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy mris across lifespan. *Medical Image Analysis*, 72:102091, 2021.

[139] T. He, R. Kong, A. J. Holmes, M. Nguyen, M. R. Sabuncu, S. B. Eickhoff, D. Bzdok, J. Feng, and B. T. Yeo. Do deep neural networks outperform kernel regression for functional connectivity prediction of behavior? *BioRxiv*, page 473603, 2018.

[140] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[141] A. Hernández-García and P. König. Do deep nets really need weight decay and dropout? *arXiv preprint arXiv:1802.07042*, 2018.

[142] A. Hernández-García, J. Mehrer, N. Kriegeskorte, P. König, and T. C. Kietzmann. Deep neural networks trained with heavier data augmentation learn features closer to representations in hit. In *Conference on Cognitive Computational Neuroscience*, volume 1, 2018.

[143] D. Hibar, L. T. Westlye, T. G. van Erp, J. Rasmussen, C. D. Leonardo, J. Faskowitz, U. K. Haukvik, C. B. Hartberg, N. T. Doan, I. Agartz, et al. Subcortical volumetric abnormalities in bipolar disorder. *Molecular psychiatry*, 21(12):1710–1716, 2016.

[144] D. Hibar, L. T. Westlye, N. T. Doan, N. Jahanshad, J. Cheung, C. R. Ching, A. Versace, A. Bilderbeck, A. Uhlmann, B. Mwangi, et al. Cortical abnormalities in bipolar disorder: an mri analysis of 6503 individuals from the enigma bipolar disorder working group. *Molecular psychiatry*, 23(4):932–942, 2018.

[145] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[146] R. Honea, T. J. Crow, D. Passingham, and C. E. Mackay. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *American Journal of Psychiatry*, 162(12):2233–2245, 2005.

[147] F. Hozer, S. Sarrazin, C. Laidi, P. Favre, M. Pauling, D. Cannon, C. McDonald, L. Emsell, J.-F. Mangin, E. Duchesnay, et al. Lithium prevents grey matter atrophy in patients with bipolar disorder: an international multicenter study. *Psychological medicine*, pages 1–10, 2020.

[148] F. Hozer, S. Sarrazin, C. Laidi, P. Favre, M. Pauling, D. Cannon, C. McDonald, L. Emsell, J.-F. Mangin, E. Duchesnay, et al. Lithium prevents grey matter atrophy in patients with bipolar disorder: an international multicenter study. *Psychological medicine*, 51(7):1201–1210, 2021.

[149] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[150] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[151] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[152] I. Hwang, E. K. Yeon, J. Y. Lee, R.-E. Yoo, K. M. Kang, T. J. Yun, S. H. Choi, C.-H. Sohn, H. Kim, and J.-h. Kim. Prediction of brain age from routine t2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network. *Neurobiology of Aging*, 105:78–85, 2021.

[153] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders, 2010.

[154] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[155] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[156] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27 (4):685–691, 2008.

[157] M. J. Jafri, G. D. Pearlson, M. Stevens, and V. D. Calhoun. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage*, 39(4):1666–1681, 2008.

[158] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[159] M. Jenkinson, M. Pechaud, S. Smith, et al. Bet2: Mr-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping*, volume 17, page 167. Toronto., 2005.

[160] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[161] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[162] E. Johnstone, C. Frith, T. Crow, J. Husband, and L. Kreel. Cerebral ventricular size and cognitive impairment in chronic schizophrenia. *The Lancet*, 308(7992):924–926, 1976.

[163] L. Jollans, R. Boyle, E. Artiges, T. Banaschewski, S. Desrivières, A. Grigis, J.-L. Martinot, T. Paus, M. N. Smolka, H. Walter, et al. Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, 199:351–365, 2019.

[164] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10 (1), Nov. 2019. doi: 10.1038/s41467-019-13163-9. URL https://doi.org/10.1038/s41467-019-13163-9.

[165] J. Kambeitz, C. Cabral, M. D. Sacchet, I. H. Gotlib, R. Zahn, M. H. Serpa, M. Walter, P. Falkai, and N. Koutsouleris. Reply to: sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. *Biological psychiatry*, 84(11): e83–e84, 2018.

[166] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[167] T. Kaufmann, D. van der Meer, N. T. Doan, E. Schwarz, M. J. Lund, I. Agartz, D. Alnæs, D. M. Barch, R. Baur-Streubel, A. Bertolino, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623, 2019.

[168] H. J. Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

[169] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[170] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[171] S. M. Kia, H. Huijsdens, R. Dinga, T. Wolfers, M. Mennes, O. A. Andreassen, L. T. Westlye, C. F. Beckmann, and A. F. Marquand. Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data, 2020.

[172] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[173] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[174] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[175] I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

[176] G. Koppe, A. Meyer-Lindenberg, and D. Durstewitz. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46(1):176–190, 2021.

[177] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[178] N. Koutsouleris, C. Davatzikos, S. Borgwardt, C. Gaser, R. Bottlender, T. Frodl, P. Falkai, A. Riecher-Rössler, H.-J. Möller, M. Reiser, et al. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophrenia bulletin*, 40(5):1140–1153, 2014.

[179] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[180] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[181] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[182] G. La Barbera, H. Boussaid, F. Maso, S. Sarnacki, L. Rouet, P. Gori, and I. Bloch. Anatomically constrained ct image translation for heterogeneous blood vessel segmentation. *BMVC*, 2022.

[183] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[184] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

[185] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. *Journal of machine learning research*, 10(1), 2009.

[186] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.

[187] Y. LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Technical report, Courant Institute of Mathematical Sciences, New York University, 2022.

[188] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[189] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[190] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[191] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

[192] P. Li and R. B. Clariana. Reading comprehension in l1 and l2: An integrative approach. *Journal of Neurolinguistics*, 50:94–105, 2019.

[193] M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, and N. Jahanshad. Style transfer using generative adversarial networks for multi-site mri harmonization. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 313–322, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

[194] S. Liu and P.-T. Yap. Learning multi-site harmonization of magnetic resonance images without traveling human phantoms, 2021.

[195] A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[196] S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

[197] A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann. Conceptualizing mental disorders as deviations from normative functioning. *Molecular psychiatry*, 24(10):1415–1424, 2019.

[198] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

[199] G. M. McAlonan, V. Cheung, C. Cheung, J. Suckling, G. Y. Lam, K. Tai, L. Yip, D. G. Murphy, and S. E. Chua. Mapping the brain in autism. a voxel-based mri study of volumetric differences and intercorrelations in autism. *Brain*, 128(2):268–276, 2005.

[200] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[201] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[202] J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

[203] L. Mottron and D. Bzdok. Autism spectrum heterogeneity: fact or artifact? *Molecular Psychiatry*, 25(12):3178–3185, 2020.

[204] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

[205] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

[206] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev, et al. Narratives: fmri data for evaluating models of naturalistic language comprehension. *bioRxiv*, pages 2020–12, 2021.

[207] B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? *arXiv preprint arXiv:2008.11687*, 2020.

[208] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[209] E. Nozari, M. A. Bertolero, J. Stiso, L. Caciagli, E. J. Cornblath, X. He, A. S. Mahadevan, G. J. Pappas, and D. S. Bassett. Is the brain macroscopically linear? a system identification of resting state dynamics. *arXiv preprint arXiv:2012.12351*, 2020.

[210] A. Nunes, H. G. Schnack, C. R. Ching, I. Agartz, T. N. Akudjedu, M. Alda, D. Alnæs, S. Alonso-Lana, J. Bauer, B. T. Baune, et al. Using structural mri to identify bipolar disorders–13 site machine learning study in 3020 individuals from the enigma bipolar disorders working group. *Molecular psychiatry*, 25(9):2130–2143, 2020.

[211] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[212] D. P. Orfanos, V. Michel, Y. Schwartz, P. Pinel, A. Moreno, D. Le Bihan, and V. Frouin. The brainomics/localizer database. *NeuroImage*, 144:309–314, 2017.

[213] Y. Østby, C. K. Tamnes, A. M. Fjell, L. T. Westlye, P. Due-Tønnessen, and K. B. Walhovd. Heterogeneity in subcortical brain development: a structural magnetic resonance imaging study of brain maturation from 8 to 30 years. *Journal of Neuroscience*, 29(38): 11772–11782, 2009.

[214] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[215] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[216] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[217] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[218] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, 2021.

[219] F. Pérez-García, R. Sparks, and S. Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv preprint arXiv:2003.04696*, 2020.

[220] M. L. Phillips and H. A. Swartz. A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of underlying neural circuitry and a road map for future research. *American Journal of Psychiatry*, 171(8):829–843, 2014.

[221] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, et al. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208: 116450, 2020.

[222] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

[223] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(2):108–120, 2019.

[224] M. Quaak, L. van de Mortel, R. M. Thomas, and G. van Wingen. Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *NeuroImage: Clinical*, 30, 2021.

[225] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[226] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[227] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[228] J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M. J. Green, C. S. Weickert, T. Weickert, J. Bruggemann, T. Kircher, et al. Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. *NeuroImage*, 218:116956, 2020.

[229] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

[230] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[231] N. Raz, F. Gunning-Dixon, D. Head, K. M. Rodrigue, A. Williamson, and J. D. Acker. Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: replicability of regional differences in volume. *Neurobiology of aging*, 25(3):377–396, 2004.

[232] L. M. Rimol, R. Nesvåg, D. J. Hagler Jr, Ø. Bergmann, C. Fennema-Notestine, C. B. Hartberg, U. K. Haukvik, E. Lange, C. J. Pung, A. Server, et al. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological psychiatry*, 71(6): 552–560, 2012.

[233] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[234] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. *International Conference on Learning Representations*, 2021.

[235] R. Robinson, Q. Dou, D. Coelho de Castro, K. Kamnitsas, M. de Groot, R. M. Summers, D. Rueckert, and B. Glocker. Image-level harmonization of multi-site data using image-and-spatial transformer networks. *Lecture Notes in Computer Science*, page 710–719, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-59728-3_69. URL http://dx.doi.org/10.1007/978-3-030-59728-3_69.

[236] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot. Automated anatomical labelling atlas 3. *Neuroimage*, 206:116189, 2020.

[237] A. F. Rosen, D. R. Roalf, K. Ruparel, J. Blake, K. Seelaus, L. P. Villa, R. Ciric, P. A. Cook, C. Davatzikos, M. A. Elliott, A. Garcia de La Garza, E. D. Gennatas, M. Quarmley, J. E. Schmitt, R. T. Shinohara, M. D. Tisdall, R. C. Craddock, R. E. Gur, R. C. Gur, and T. D. Satterthwaite. Quantitative assessment of structural image quality. *NeuroImage*, 169:407–418, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.12.059. URL https://www.sciencedirect.com/science/article/pii/S1053811917310832.

[238] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[239] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[240] R. Salvador, J. Radua, E. J. Canales-Rodríguez, A. Solanes, S. Sarró, J. M. Goikolea, A. Valiente, G. C. Monté, M. d. C. Natividad, A. Guerrero-Pedraza, et al. Evaluation of machine learning algorithms and structural features for optimal mri-based diagnostic prediction in psychosis. *PLoS One*, 12(4):e0175683, 2017.

[241] S. Sarrazin, A. Cachia, F. Hozer, C. McDonald, L. Emsell, D. M. Cannon, M. Wessa, J. Linke, A. Versace, N. Hamdani, et al. Neurodevelopmental subtypes of bipolar disorder are related to cortical folding patterns: An international multicenter study. *Bipolar disorders*, 20(8):721–732, 2018.

[242] D. Sasabayashi, T. Takahashi, Y. Takayanagi, and M. Suzuki. Anomalous brain gyrification patterns in major psychiatric disorders: a systematic review and transdiagnostic integration. *Translational psychiatry*, 11(1):1–12, 2021.

[243] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019.

[244] N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

[245] D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[246] H. G. Schnack and R. S. Kahn. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Frontiers in psychiatry*, 7:50, 2016.

[247] H. G. Schnack, N. E. Van Haren, M. Nieuwenhuis, H. E. Hulshoff Pol, W. Cahn, and R. S. Kahn. Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. *American Journal of Psychiatry*, 173(6):607–616, 2016.

[248] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. doi: 10.1109/CVPR.2015.7298682. URL http://arxiv.org/abs/1503.03832. arXiv: 1503.03832.

[249] M.-A. Schulz, B. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok. Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nature communications*, 11 (1):1–15, 2020.

[250] M.-A. Schulz, D. Bzdok, S. Haufe, J.-D. Haynes, and K. Ritter. Performance reserves in brain-imaging-based phenotype prediction. *bioRxiv*, 2022.

[251] E. Schwarz, N. T. Doan, G. Pergola, L. T. Westlye, T. Kaufmann, T. Wolfers, R. Brecheisen, T. Quarto, A. J. Ing, P. Di Carlo, et al. Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder. *Translational psychiatry*, 9(1):1–13, 2019.

[252] K. Sechidis, G. Tsoumakas, and I. Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.

[253] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

[254] R. Shaw, C. H. Sudre, S. Ourselin, and M. J. Cardoso. Mri k-space motion artefact augmentation: Model robustness and task-specific uncertainty. In *MIDL*, pages 427–436, 2019.

[255] D. Shen and C. Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging*, 21(11):1421–1439, 2002.

[256] M. E. Shenton, C. C. Dickey, M. Frumin, and R. W. McCarley. A review of mri findings in schizophrenia. *Schizophrenia research*, 49(1-2):1–52, 2001.

[257] G. Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

[258] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[259] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[260] P. Smolensky, R. T. McCoy, R. Fernandez, M. Goldrick, and J. Gao. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems. *AI Magazine*, 2022.

[261] K. Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.nips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

[262] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[263] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[264] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.

[265] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[266] D. Steinkraus, I. Buck, and P. Simard. Using gpus for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1115–1120. IEEE, 2005.

[267] J. Sui, S. Qi, T. G. van Erp, J. Bustillo, R. Jiang, D. Lin, J. A. Turner, E. Damaraju, A. R. Mayer, Y. Cui, et al. Multimodal neuromarkers in schizophrenia via cognition-guided mri fusion. *Nature communications*, 9(1):1–14, 2018.

[268] K. Sukel. Neuroanatomy: The basics, 2019. URL https://www.dana.org/article/neuroanatomy-the-basics/.

[269] A. Sunavsky and J. Poppenk. Neuroimaging predictors of creativity in healthy adults. *Neuroimage*, 206:116292, 2020.

[270] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert. 3d self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems*, volume 33, pages 18158–18172, 2020.

[271] C. A. Tamminga, G. Pearlson, M. Keshavan, J. Sweeney, B. Clementz, and G. Thaker. Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. *Schizophrenia bulletin*, 40:S131–S137, 2014.

[272] X. Tao, Y. Li, W. Zhou, K. Ma, and Y. Zheng. Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 238–248. Springer, 2020.

[273] E. Tartaglione, C. A. Barbano, and M. Grangetto. End: Entangling and disentangling deep representations for bias correction. *arXiv preprint arXiv:2103.02023*, 2021.

[274] D. Teney, E. Abbasnejad, S. Lucey, and A. van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022.

[275] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[276] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839, 2020.

[277] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[278] M. E. Torbati, D. L. Tudorascu, D. S. Minhas, P. Maillard, C. S. DeCarli, and S. J. Hwang. Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3284–3293, October 2021.

[279] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[280] Y.-H. H. Tsai, T. Li, M. Q. Ma, H. Zhao, K. Zhang, L.-P. Morency, and R. Salakhutdinov. Conditional contrastive learning with kernel. *arXiv preprint arXiv:2202.05458*, 2022.

[281] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[282] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6): 1310–1320, 2010.

[283] J. M. Valverde, V. Imani, A. Abdollahzadeh, R. De Feo, M. Prakash, R. Ciszek, and J. Tohka. Transfer learning in magnetic resonance brain imaging: a systematic review. *Journal of imaging*, 7(4):66, 2021.

[284] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[285] T. G. van Erp, D. P. Hibar, J. M. Rasmussen, D. C. Glahn, G. D. Pearlson, O. A. Andreassen, I. Agartz, L. T. Westlye, U. K. Haukvik, A. M. Dale, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium. *Molecular psychiatry*, 21(4):547–553, 2016.

[286] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[287] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of mr images of the brain. *IEEE transactions on medical imaging*, 18 (10):897–908, 1999.

[288] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[289] G. Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180:68–77, 2018.

[290] G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8, 2022.

[291] S. Vieira, Q.-y. Gong, W. H. Pinaya, C. Scarpazza, S. Tognin, B. Crespo-Facorro, D. Tordesillas-Gutierrez, V. Ortiz-García, E. Setien-Suero, F. E. Scheepers, et al. Using

machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. *Schizophrenia bulletin*, 46(1):17–26, 2020.

[292] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[293] C. Wachinger, A. Rieckmann, S. Pölsterl, A. D. N. Initiative, et al. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, 2021.

[294] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

[295] J. Wang, Y. song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning Fine-grained Image Similarity with Deep Ranking. In *CVPR*, 2014.

[296] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

[297] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[298] X. Wang, Y. Hua, E. Kodirov, and N. M. Robertson. Ranked List Loss for Deep Metric Learning. In *CVPR*, 2019.

[299] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022.

[300] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

[301] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[302] C. Wei, H. Wang, W. Shen, and A. Yuille. {CO}2: Consistent contrast for unsupervised visual representation learning. In *International Conference on Learning Representations*, 2021.

[303] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. Technical report, California Institute of Technology, 2010.

[304] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, et al. Convolutional neural networks

for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, page 101694, 2020.

[305] T. Wolfers, N. T. Doan, T. Kaufmann, D. Alnæs, T. Moberget, I. Agartz, J. K. Buitelaar, T. Ueland, I. Melle, B. Franke, et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA psychiatry*, 75(11): 1146–1155, 2018.

[306] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, 20(3):365–377, 2017.

[307] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[308] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[309] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun. Decoupled contrastive learning. *arXiv preprint arXiv:2110.06848*, 2021.

[310] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf.

[311] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[312] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[313] M. Zabihi, D. L. Floris, S. M. Kia, T. Wolfers, J. Tillmann, A. L. Arenas, C. Moessnang, T. Banaschewski, R. Holt, S. Baron-Cohen, et al. Fractionating autism based on neuroanatomical normative modeling. *Translational psychiatry*, 10(1):1–10, 2020.

[314] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[315] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[316] C. Zhang, M.-H. Hsieh, and D. Tao. Generalization bounds for vicinal risk minimization principle. *arXiv preprint arXiv:1811.04351*, 2018.

[317] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[318] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[319] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[320] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021.

[321] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–428. Springer, 2019.

[322] J. Zhuo and R. P. Gullapalli. Mr artifacts, safety, and quality control. *Radiographics*, 26 (1):275–297, 2006.

[323] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

[324] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1 (1):1–13, 2014.